

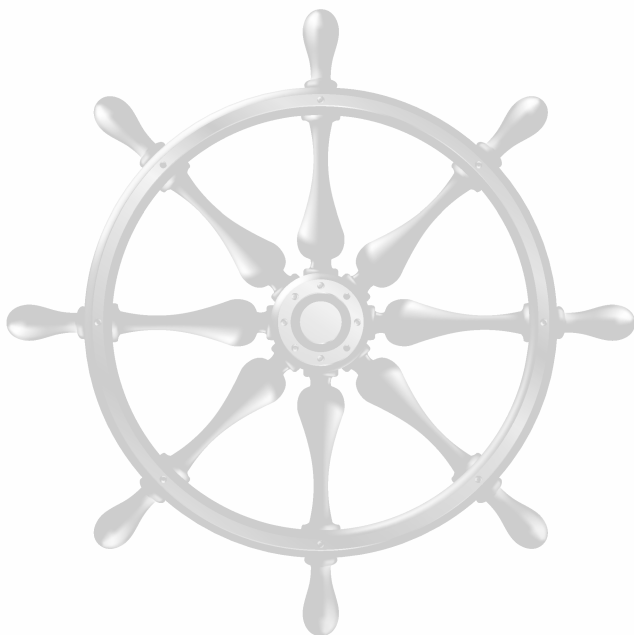
# Kubernetes

## 权威指南



从Docker到Kubernetes  
实践全接触

龚正 吴治辉 王伟 / 等编著  
崔秀龙 闫健勇



电子工业出版社  
Publishing House of Electronics Industry  
北京•BEIJING

## 内 容 简 介

Kubernetes 是由谷歌开源的 Docker 容器集群管理系统，为容器化的应用提供了资源调度、部署运行、服务发现、扩容及缩容等一整套功能。《Kubernetes 权威指南：从 Docker 到 Kubernetes 实践全接触（纪念版）》从架构师、开发人员和运维人员的角度，阐述了 Kubernetes 的基本概念、实践指南、核心原理、开发指导、运维指南及源码分析等内容，图文并茂、内容丰富、由浅入深、讲解全面；围绕着生产环境中可能出现的问题，给出了大量的典型案例，比如安全配置、网络方案、共享存储方案、高可用性方案及 Trouble Shooting 技巧等，有很强的实战指导意义。本书随着 Kubernetes 版本更新不断完善，目前涵盖了 Kubernetes 从 v1.0 到 v1.6 版本的全部特性，尽力为 Kubernetes 用户提供全方位的指南。

无论是对于软件工程师、测试工程师、运维工程师、软件架构师、技术经理，还是对于资深 IT 人士来说，本书都极具参考价值。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究。

## 图书在版编目（CIP）数据

Kubernetes 权威指南：从 Docker 到 Kubernetes 实践全接触：纪念版 / 龚正等编著. —北京：电子工业出版社，2017.9

ISBN 978-7-121-32351-5

I. ①K… II. ①龚… III. ①Linux 操作系统—程序设计—指南 IV. ①TP316.85-62

中国版本图书馆 CIP 数据核字（2017）第 182028 号

策划编辑：张国霞

责任编辑：徐津平

印 刷：

装 订：

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱 邮编 100036

开 本：787×980 1/16 印张：43.25 字数：970 千字

版 次：2017 年 9 月第 1 版

印 次：2017 年 9 月第 1 次印刷

印 数：2500 册 定价：119.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 [zlt@phei.com.cn](mailto:zlt@phei.com.cn)，盗版侵权举报请发邮件至 [dbqq@phei.com.cn](mailto:dbqq@phei.com.cn)。

本书咨询联系方式：010-51260888-819，[faq@phei.com.cn](mailto:faq@phei.com.cn)。

# 推荐序

经过作者们多年的实践经验积累及近一年的精心准备，本书终于与我们大家见面了。我有幸作为首批读者，提前见证和学习了在云时代引领业界技术方向的 **Kubernetes** 和 **Docker** 的最新动态。

从内容上讲，本书从一个开发者的角度去理解、分析和解决问题：从基础入门到架构原理，从运行机制到开发源码，再从系统运维到应用实践，讲解全面。本书图文并茂，内容丰富，由浅入深，对基本原理阐述清晰，对程序源码分析透彻，对实践经验体会深刻。

我认为本书值得推荐的原因有以下几点。

首先，作者的所有观点和经验，均是在多年建设、维护大型应用系统的过程中积累形成的。例如，读者通过学习书中的 **Kubernetes** 运维指南和高级应用实践案例章节的内容，不仅可以直接提高开发技能，还可以解决在实践过程中经常遇到的各种关键问题。书中的这些内容具有很高的借鉴和推广意义。

其次，通过大量的实例操作和详尽的源码解析，本书可以帮助读者进一步深刻理解 **Kubernetes** 的各种概念。例如书中“**Java** 访问 **Kubernetes API**”的几种方法，读者参照其中的案例，只要稍做修改，再结合实际的应用需求，就可以用于正在开发的项目中，达到事半功倍的效果，对于有一定 **Java** 基础的专业人士快速学习 **Kubernetes** 的各种细节和实践操作十分有利。

再次，为了让初学者快速入门，本书配备了即时在线交流工具和专业后台技术支持团队。如果你在开发和应用的过程中遇到各类相关问题，均可直接联系该团队的开发支持专家。

最后，我们可以看到，容器化技术已经成为计算模型演化的一个开端，**Kubernetes** 作为谷歌开源的 **Docker** 容器集群管理技术，在这场新的技术革命中扮演着重要的角色。**Kubernetes** 正在被众多知名企业所采用，例如 **RedHat**、**VMware**、**CoreOS** 及腾讯等，因此，**Kubernetes** 站在了容器新技术变革的浪潮之巅，将具有不可预估的发展前景和商业价值。

## Kubernetes 权威指南：从 Docker 到 Kubernetes 实践全接触（纪念版）

如果你是初级程序员，那么你有必要好好学习本书；如果你正在 IT 领域进行高级进阶修炼，那你也有必要阅读本书。无论是架构师、开发者、运维人员，还是对容器技术好奇的读者，本书都是一本不可多得的带你从入门向高级进阶的精品书，值得大家选择！

初瑞

中国移动业务支撑中心高级经理





# 自序

我不知道你是如何获得这本书的，可能是在百度头条、网络广告、朋友圈中听说本书后购买的，也可能是某一天逛书店时，这本书恰好神奇地出现在你面前的书架上，让你想起一千多年前那个意外得到《太公兵法》的传奇少年，你觉得这是冥冥之中上天的恩赐，于是果断带走。不管怎样，我相信多年以后，这本书仍然值得你回忆。

Kubernetes 这个名字起源于古希腊，是舵手的意思，所以它的 Logo 既像一张渔网，又像一个罗盘。谷歌采用这个名字的一层深意就是：既然 Docker 把自己定位为驮着集装箱在大海上自在遨游的鲸鱼，那么谷歌就要以 Kubernetes 掌舵大航海时代的话语权，“捕获”和“指引”这条鲸鱼按照“主人”设定的路线巡游，确保谷歌倾力打造的新一代容器世界的宏伟蓝图顺利实现。

虽然 Kubernetes 自诞生至今才 1 年多，其第一个正式版本 Kubernetes 1.0 于 2015 年 7 月才发布，完全是个新生事物，但其影响力巨大，已经吸引了包括 IBM、惠普、微软、红帽、Intel、VMware、CoreOS、Docker、Mesosphere、Mirantis 等在内的众多业界巨头纷纷加入。红帽这个软件虚拟化领域的领导者之一，在容器技术方面已经完全“跟从”谷歌了，不仅把自家的第三代 OpenShift 产品的架构底层换成了 Docker+Kubernetes，还直接在其新一代容器操作系统 Atomic 内原生集成了 Kubernetes。

Kubernetes 是第一个将“一切以服务（Service）为中心，一切围绕服务运转”作为指导思想创新型产品，它的功能和架构设计自始至终地遵循了这一指导思想，构建在 Kubernetes 上的系统不仅可以独立运行在物理机、虚拟机集群或者企业私有云上，也可以被托管在公有云中。Kubernetes 方案的另一个亮点是自动化，在 Kubernetes 的解决方案中，一个服务可以自我扩展、自我诊断，并且容易升级，在收到服务扩容的请求后，Kubernetes 会触发调度流程，最终在选定的目标节点上启动相应数量的服务实例副本，这些副本在启动成功后会自动加入负载均衡器中并生效，整个过程无须额外的人工操作。另外，Kubernetes 会定时巡查每个服务的所有实例的可用性，确保服务实例的数量始终保持为预期的数量，当它发现某个实例不可用时，会自动

重启该实例或者在其他节点重新调度、运行一个新实例，这样，一个复杂的过程无须人工干预即可全部自动化完成。试想一下，如果一个包括几十个节点且运行着几万个容器的复杂系统，其负载均衡、故障检测和故障修复等都需要人工介入进行处理，那将是多么的难以想象。

通常我们会把 Kubernetes 看作 Docker 的上层架构，就好像 Java 与 J2EE 的关系一样：J2EE 是以 Java 为基础的企业级软件架构，而 Kubernetes 则以 Docker 为基础打造了一个云计算时代的全新分布式系统架构。但 Kubernetes 与 Docker 之间还存在着更为复杂的关系，从表面上看，似乎 Kubernetes 离不开 Docker，但实际上在 Kubernetes 的架构里，Docker 只是其目前支持的两种底层容器技术之一，另一个容器技术则是 Rocket，后者来源于 CoreOS 这个 Docker 昔日的“恋人”所推出的竞争产品。

Kubernetes 同时支持这两种互相竞争的容器技术，这是有深刻的历史原因的。快速发展的 Docker 打败了谷歌曾经名噪一时的开源容器技术 lsmctfy，并迅速风靡世界。但是，作为一个已经对全球 IT 公司产生重要影响的技术，Docker 背后的容器标准的制定注定不可能被任何一个公司私有控制，于是就有了后来引发危机的 CoreOS 与 Docker 分手事件，其导火索是 CoreOS 撇开了 Docker，推出了与 Docker 相对抗的开源容器项目——Rocket，并动员一些知名 IT 公司成立委员会来试图主导容器技术的标准化，该分手事件愈演愈烈，最终导致 CoreOS “傍上”谷歌一起宣布“叛逃”Docker 阵营，共同发起了基于 CoreOS+Rocket+Kubernetes 的新项目 Tectonic。这让当时的 Docker 阵营和 Docker 粉丝们无比担心 Docker 的命运，不管最终鹿死谁手，容器技术分裂态势的加剧对所有牵涉其中的人来说都没有好处，于是 Linux 基金会出面调和矛盾，双方都退让一步，最终的结果是 Linux 基金会于 2015 年 6 月宣布成立开放容器技术项目（Open Container Project），谷歌、CoreOS 及 Docker 都加入了 OCP 项目。但通过查看 OCP 项目的成员名单，你会发现 Docker 在这个名单中只能算一个小角色了。OCP 的成立最终结束了这场让无数人揪心的“战争”，Docker 公司被迫放弃了自己的独家控制权。作为回报，Docker 的容器格式被 OCP 采纳为新标准的基础，并且由 Docker 负责起草 OCP 草案规范的初稿文档，当然这个“标准起草者”的角色也不是那么容易担当的，Docker 要提交自己的容器执行引擎的源码作为 OCP 项目的启动资源。

事到如今，我们再来回顾当初 CoreOS 与谷歌的叛逃事件，从表面上看，谷歌貌似是被诱拐“出柜”的，但局里人都明白，谷歌才是这一系列事件背后的主谋，其不仅为当年失败的 lsmctfy 报了一箭之仇，还重新掌控了容器技术的未来。容器标准之战大捷之后，谷歌进一步扩大了联盟并提高了自身影响力。2015 年 7 月，谷歌正式宣布加入 OpenStack 阵营，其目标是确保 Linux 容器及关联的容器管理技术 Kubernetes 能够被 OpenStack 生态圈所容纳，并且成为 OpenStack 平台上与 KVM 虚拟机一样的一等公民。谷歌加入 OpenStack 意味着对数据中心控制平面的争夺已经结束，以容器为代表的形态与以虚拟化为代表的系统形态将会完美融合于 OpenStack 之上，并与软件定义网络 and 软件定义存储一起统治下一代数据中心。

谷歌凭借着几十年大规模容器使用的丰富经验，步步为营，先是祭出 Kubernetes 这个神器，然后又掌控了容器技术的制定标准，最后又入驻 OpenStack 阵营全力将 Kubernetes 扶上位，谷歌这个 IT 界的领导者和创新者再次王者归来。我们都明白，在 IT 世界里只有那些被大公司掌控和推广的，同时被业界众多巨头都认可和支持的新技术才能生存和壮大下去。Kubernetes 就是当今 IT 界里符合要求且为数不多的热门技术之一，它的影响力可能长达十年，所以，我们每个 IT 人都有理由重视这门新技术。

谁能比别人领先一步掌握新技术，谁就在竞争中赢得了先机。惠普中国电信解决方案领域的资深专家团一起分工协作，并行研究，废寝忘食地合力撰写，完成了这部近 700 页的 Kubernetes 权威指南。经过两年的高速发展，Kubernetes 先后发布了 v1.0~v1.6 这 6 个大版本，每个版本都带来了大量的新特性，能够处理的应用场景也越来越丰富。本书遵循从入门到精通的学习路线，全书共分为六大章节，涵盖了入门、实践指南、架构原理、开发指南、高级案例、运维指南和源码分析等内容，内容详实、图文并茂，几乎囊括了 Kubernetes 到 v1.6 版本的方方面面，无论是对于软件工程师、测试工程师、运维工程师、软件架构师、技术经理，还是对于资深 IT 人士来说，本书都极具参考价值。

吴治辉

惠普公司系统架构师

轻松注册成为博文视点社区用户（[www.broadview.com.cn](http://www.broadview.com.cn)），您即可享受以下服务。

- **提交勘误：**您对书中内容的修改意见可在【提交勘误】处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动：**在页面下方【读者评论】处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：<http://www.broadview.com.cn/32351>

二维码：



# 目 录

## 第 1 章 Kubernetes 入门

1

1.1	Kubernetes 是什么	1
1.2	为什么要用 Kubernetes	4
1.3	从一个简单的例子开始	5
1.3.1	环境准备	6
1.3.2	启动 MySQL 服务	6
1.3.3	启动 Tomcat 应用	9
1.3.4	通过浏览器访问网页	10
1.4	Kubernetes 基本概念和术语	12
1.4.1	Master	12
1.4.2	Node	12
1.4.3	Pod	15
1.4.4	Label (标签)	18
1.4.5	Replication Controller	22
1.4.6	Deployment	26
1.4.7	Horizontal Pod Autoscaler	28
1.4.8	StatefulSet	29
1.4.9	Service (服务)	30

1.4.10	Volume（存储卷）	37
1.4.11	Persistent Volume	41
1.4.12	Namespace（命名空间）	42
1.4.13	Annotation（注解）	43
1.4.14	小结	44
<b>第 2 章</b>	<b>Kubernetes 实践指南</b>	<b>45</b>
2.1	Kubernetes 安装与配置	45
2.1.1	系统要求	45
2.1.2	使用 kubeadm 工具快速安装 Kubernetes 集群	46
2.1.3	以二进制文件方式安装 Kubernetes 集群	51
2.1.4	Kubernetes 集群的安全设置	59
2.1.5	Kubernetes 集群的网络配置	64
2.1.6	内网中的 Kubernetes 相关配置	64
2.1.7	Kubernetes 的版本升级	65
2.1.8	Kubernetes 核心服务配置详解	66
2.2	kubectl 命令行工具用法详解	86
2.2.1	kubectl 用法概述	86
2.2.2	kubectl 子命令详解	88
2.2.3	kubectl 参数列表	90
2.2.4	kubectl 输出格式	90
2.2.5	kubectl 操作示例	92
2.3	深入掌握 Pod	93
2.3.1	Pod 定义详解	93
2.3.2	Pod 的基本用法	98
2.3.3	静态 Pod	103

2.3.4	Pod 容器共享 Volume .....	104
2.3.5	Pod 的配置管理 .....	106
2.3.6	在容器内获取 Pod 信息（Downward API） .....	119
2.3.7	Pod 生命周期和重启策略 .....	124
2.3.8	Pod 健康检查 .....	125
2.3.9	玩转 Pod 调度 .....	127
2.3.10	Init Container（初始化容器） .....	149
2.3.11	Pod 的升级和回滚 .....	152
2.3.12	Pod 的扩容和缩容 .....	166
2.3.13	使用 StatefulSet 搭建 MongoDB 集群 .....	171
2.4	深入掌握 Service .....	180
2.4.1	Service 定义详解 .....	181
2.4.2	Service 基本用法 .....	182
2.4.3	Headless Service .....	187
2.4.4	集群外部访问 Pod 或 Service .....	192
2.4.5	DNS 服务搭建指南 .....	196
2.4.6	自定义 DNS 和上游 DNS 服务器 .....	204
2.4.7	Ingress：HTTP 7 层路由机制 .....	208

第3章 Kubernetes 核心原理

226

3.1	Kubernetes API Server 原理分析 .....	226
3.1.1	Kubernetes API Server 概述 .....	226
3.1.2	独特的 Kubernetes Proxy API 接口 .....	229
3.1.3	集群功能模块之间的通信 .....	230
3.2	Controller Manager 原理分析 .....	231
3.2.1	Replication Controller .....	232

3.2.2	Node Controller	234
3.2.3	ResourceQuota Controller	235
3.2.4	Namespace Controller	237
3.2.5	Service Controller 与 Endpoint Controller	237
3.3	Scheduler 原理分析	238
3.4	kubelet 运行机制分析	242
3.4.1	节点管理	242
3.4.2	Pod 管理	243
3.4.3	容器健康检查	244
3.4.4	cAdvisor 资源监控	245
3.5	kube-proxy 运行机制分析	247
3.6	深入分析集群安全机制	251
3.6.1	API Server 认证管理 (Authentication)	251
3.6.2	API Server 授权管理 (Authorization)	253
3.6.3	Admission Control (准入控制)	272
3.6.4	Service Account	274
3.6.5	Secret 私密凭据	279
3.7	网络原理	282
3.7.1	Kubernetes 网络模型	282
3.7.2	Docker 的网络基础	284
3.7.3	Docker 的网络实现	296
3.7.4	Kubernetes 的网络实现	304
3.7.5	Pod 和 Service 网络实战	308
3.7.6	CNI 网络模型	321
3.7.7	Kubernetes 网络策略	331
3.7.8	开源的网络组件	333

3.8 共享存储原理	363
3.8.1 共享存储机制概述	363
3.8.2 PV 详解	364
3.8.3 PVC 详解	368
3.8.4 PV 和 PVC 的生命周期	370
3.8.5 StorageClass 详解	373
3.8.6 动态存储管理实战：GlusterFS	376

第4章 Kubernetes 开发指南388

---

4.1 REST 简述	388
4.2 Kubernetes API 详解	390
4.2.1 Kubernetes API 概述	390
4.2.2 API 版本	395
4.2.3 API Groups（API 组）	395
4.2.4 API 方法说明	397
4.2.5 API 响应说明	398
4.3 使用 Java 程序访问 Kubernetes API	400
4.3.1 Jersey	401
4.3.2 Fabric8	412
4.3.3 使用说明	413

第5章 Kubernetes 运维指南434

---

5.1 Kubernetes 集群管理指南	434
5.1.1 Node 的管理	434
5.1.2 更新资源对象的 Label	436
5.1.3 Namespace：集群环境共享与隔离	437



5.1.4	Kubernetes 资源管理	441
5.1.5	资源紧缺时的 Pod 驱逐机制	475
5.1.6	Pod Disruption Budget (主动驱逐保护)	483
5.1.7	Kubernetes 集群的高可用部署方案	485
5.1.8	Kubernetes 集群监控	496
5.1.9	集群统一日志管理	513
5.1.10	Kubernetes 审计日志 (Audit Log)	522
5.1.11	使用 Web UI (Dashboard) 管理集群	523
5.1.12	Helm: Kubernetes 应用包管理工具	527
5.2	Trouble Shooting 指导	538
5.2.1	查看系统 Event 事件	538
5.2.2	查看容器日志	540
5.2.3	查看 Kubernetes 服务日志	541
5.2.4	常见问题	542
5.2.5	寻求帮助	546
5.3	Kubernetes 开发中的新功能	546
5.3.1	Pod Preset (运行时参数注入策略)	546
5.3.2	Cluster Federation (集群联邦)	553
5.3.3	容器运行时接口 (Container Runtime Interface-CRI)	557
5.3.4	对 GPU 的支持	561
5.3.5	Kubernetes 的演进路线 (Roadmap) 和开发模式	565

## 第 6 章 Kubernetes 源码导读

568

6.1	Kubernetes 源码结构和编译步骤	568
6.2	kube-apiserver 进程源码分析	572
6.2.1	进程启动过程	572

6.2.2	关键代码分析	574
6.2.3	设计总结	589
6.3	kube-controller-manager 进程源码分析	592
6.3.1	进程启动过程	592
6.3.2	关键代码分析	595
6.3.3	设计总结	603
6.4	kube-scheduler 进程源码分析	605
6.4.1	进程启动过程	605
6.4.2	关键代码分析	610
6.4.3	设计总结	617
6.5	kubelet 进程源码分析	619
6.5.1	进程启动过程	619
6.5.2	关键代码分析	624
6.5.3	设计总结	647
6.6	kube-proxy 进程源码分析	648
6.6.1	进程启动过程	648
6.6.2	关键代码分析	650
6.6.3	设计总结	665
6.7	kubectl 进程源码分析	666
6.7.1	kubectl create 命令	667
6.7.2	rolling-update 命令	671

# 第 1 章

# Kubernetes 入门

---

## 1.1 Kubernetes 是什么

---

Kubernetes 是什么？

首先，它是一个全新的基于容器技术的分布式架构领先方案。这个方案虽然还很新，但它是谷歌十几年以来大规模应用容器技术的经验积累和升华的一个重要成果。确切地说，Kubernetes 是谷歌严格保密十几年的秘密武器——Borg 的一个开源版本。Borg 是谷歌的一个久负盛名的内部使用的大规模集群管理系统，它基于容器技术，目的是实现资源管理的自动化，以及跨多个数据中心的资源利用率的最大化。十几年来，谷歌一直通过 Borg 系统管理着数量庞大的应用程序集群。由于谷歌员工都签署了保密协议，即便离职也不能泄露 Borg 的内部设计，所以外界一直无法了解关于它的更多信息。直到 2015 年 4 月，传闻许久的 Borg 论文伴随 Kubernetes 的高调宣传被谷歌首次公开，大家才得以了解它的更多内幕。正是由于站在 Borg 这个前辈的肩膀上，吸取了 Borg 过去十年间的经验与教训，所以 Kubernetes 一经开源就一鸣惊人，并迅速称霸了容器技术领域。

其次，如果我们的系统设计遵循了 Kubernetes 的设计思想，那么传统系统架构中那些和业务没有多大关系的底层代码或功能模块，都可以立刻从我们的视线中消失，我们不必再费心于负载均衡器的选型和部署实施问题，不必再考虑引入或自己开发一个复杂的服务治理框架，不必再头疼于服务监控和故障处理模块的开发。总之，使用 Kubernetes 提供的解决方案，我们不仅节省了不少于 30% 的开发成本，同时可以将精力更加集中于业务本身，而且由于 Kubernetes 提供了强大的自动化机制，所以系统后期的运维难度和运维成本大幅度降低。

然后，Kubernetes 是一个开放的开发平台。与 J2EE 不同，它不局限于任何一种语言，没有

限定任何编程接口，所以不论是用 Java、Go、C++还是用 Python 编写的服务，都可以毫无困难地映射为 Kubernetes 的 Service，并通过标准的 TCP 通信协议进行交互。此外，由于 Kubernetes 平台对现有的编程语言、编程框架、中间件没有任何侵入性，因此现有的系统也很容易改造升级并迁移到 Kubernetes 平台上。

最后，Kubernetes 是一个完备的分布式系统支撑平台。Kubernetes 具有完备的集群管理能力，包括多层次的安全防护和准入机制、多租户应用支撑能力、透明的服务注册和服务发现机制、内建智能负载均衡器、强大的故障发现和自我修复能力、服务滚动升级和在线扩容能力、可扩展的资源自动调度机制，以及多粒度的资源配额管理能力。同时，Kubernetes 提供了完善的管理工具，这些工具涵盖了包括开发、部署测试、运维监控在内的各个环节。因此，Kubernetes 是一个全新的基于容器技术的分布式架构解决方案，并且是一个一站式的完备的分布式系统开发和支撑平台。

在正式开始本章的 Hello World 之旅之前，我们首先要学习 Kubernetes 的一些基本知识，这样我们才能理解 Kubernetes 提供的解决方案。

在 Kubernetes 中，Service（服务）是分布式集群架构的核心，一个 Service 对象拥有如下关键特征。

- ◎ 拥有一个唯一指定的名字（比如 mysql-server）。
- ◎ 拥有一个虚拟 IP（Cluster IP、Service IP 或 VIP）和端口号。
- ◎ 能够提供某种远程服务能力。
- ◎ 被映射到了提供这种服务能力的一组容器应用上。

Service 的服务进程目前都基于 Socket 通信方式对外提供服务，比如 Redis、Memcache、MySQL、Web Server，或者是实现了某个具体业务的一个特定的 TCP Server 进程。虽然一个 Service 通常由多个相关的服务进程来提供服务，每个服务进程都有一个独立的 Endpoint（IP+Port）访问点，但 Kubernetes 能够让我们通过 Service（虚拟 Cluster IP +Service Port）连接到指定的 Service 上。有了 Kubernetes 内建的透明负载均衡和故障恢复机制，不管后端有多少服务进程，也不管某个服务进程是否会由于发生故障而重新部署到其他机器，都不会影响到我们对服务的正常调用。更重要的是这个 Service 本身一旦创建就不再变化，这意味着在 Kubernetes 集群中，我们再也不用为了服务的 IP 地址变来变去的问题而头疼了。

容器提供了强大的隔离功能，所以有必要把为 Service 提供服务的这组进程放入容器中进行隔离。为此，Kubernetes 设计了 Pod 对象，将每个服务进程包装到相应的 Pod 中，使其成为 Pod 中运行的一个容器（Container）。为了建立 Service 和 Pod 间的关联关系，Kubernetes 首先给每个 Pod 贴上一个标签（Label），给运行 MySQL 的 Pod 贴上 name=mysql 标签，给运行 PHP 的

Pod 贴上 `name=php` 标签，然后给相应的 Service 定义标签选择器 (Label Selector)，比如 MySQL Service 的标签选择器的选择条件为 `name=mysql`，意为该 Service 要作用于所有包含 `name=mysql` Label 的 Pod 上。这样一来，就巧妙地解决了 Service 与 Pod 的关联问题。

说到 Pod，我们这里先简单介绍其概念。首先，Pod 运行在一个我们称之为节点 (Node) 的环境中，这个节点既可以是物理机，也可以是私有云或者公有云中的一个虚拟机，通常在一个节点上运行几百个 Pod；其次，每个 Pod 里运行着一个特殊的被称之为 Pause 的容器，其他容器则为业务容器，这些业务容器共享 Pause 容器的网络栈和 Volume 挂载卷，因此它们之间的通信和数据交换更为高效，在设计时我们可以充分利用这一特性将一组密切相关的服务进程放入同一个 Pod 中；最后，需要注意的是，并不是每个 Pod 和它里面运行的容器都能“映射”到一个 Service 上，只有那些提供服务（无论是对内还是对外）的一组 Pod 才会被“映射”成一个服务。

在集群管理方面，Kubernetes 将集群中的机器划分为一个 Master 节点和一群工作节点(Node)。其中，在 Master 节点上运行着集群管理相关的一组进程 `kube-apiserver`、`kube-controller-manager` 和 `kube-scheduler`，这些进程实现了整个集群的资源管理、Pod 调度、弹性伸缩、安全控制、系统监控和纠错等管理功能，并且都是全自动完成的。Node 作为集群中的工作节点，运行真正的应用程序，在 Node 上 Kubernetes 管理的最小运行单元是 Pod。Node 上运行着 Kubernetes 的 `kubelet`、`kube-proxy` 服务进程，这些服务进程负责 Pod 的创建、启动、监控、重启、销毁，以及实现软件模式的负载均衡器。

最后，我们再来看看传统的 IT 系统中服务扩容和服务升级这两个难题，以及 Kubernetes 所提供的全新解决思路。服务的扩容涉及资源分配（选择哪个节点进行扩容）、实例部署和启动等环节，在一个复杂的业务系统中，这两个问题基本上靠人工一步步操作才得以完成，费时费力又难以保证实施质量。

在 Kubernetes 集群中，你只需为需要扩容的 Service 关联的 Pod 创建一个 RC (Replication Controller)，则该 Service 的扩容以至于后来的 Service 升级等头疼问题都迎刃而解。在一个 RC 定义文件中包括以下 3 个关键信息。

- ◎ 目标 Pod 的定义。
- ◎ 目标 Pod 需要运行的副本数量 (Replicas)。
- ◎ 要监控的目标 Pod 的标签 (Label)。

在创建好 RC（系统将自动创建好 Pod）后，Kubernetes 会通过 RC 中定义的 Label 筛选出对应的 Pod 实例并实时监控其状态和数量，如果实例数量少于定义的副本数量 (Replicas)，则会根据 RC 中定义的 Pod 模板来创建一个新的 Pod，然后将此 Pod 调度到合适的 Node 上启动运行，直到 Pod 实例的数量达到预定目标。这个过程完全是自动化的，无须人工干预。有了 RC，

服务的扩容就变成了一个纯粹的简单数字游戏了，只要修改 RC 中的副本数量即可。后续的 Service 升级也将通过修改 RC 来自动完成。

以将在第 2 章介绍的 PHP+Redis 留言板应用为例，只要为 PHP 留言板程序（frontend）创建一个有 3 个副本的 RC+Service，为 Redis 读写分离集群创建两个 RC：写节点（redis-master）创建一个单副本的 RC+Service，读节点（redis-slaver）创建一个有两个副本的 RC+Service，就可以分分钟完成整个集群的搭建过程了，是不是很简单？

## 1.2 为什么要用 Kubernetes

使用 Kubernetes 的理由很多，最根本的一个理由就是：IT 从来都是一个由新技术驱动的行业。

Docker 这个新兴的容器化技术当前已经被很多公司所采用，其从单机走向集群已成为必然，而云计算的蓬勃发展正在加速这一进程。Kubernetes 作为当前唯一被业界广泛认可和看好的 Docker 分布式系统解决方案，可以预见，在未来几年内，会有大量的新系统选择它，不管这些系统是运行在企业本地服务器上还是被托管到公有云上。

使用了 Kubernetes 又会收获哪些好处呢？

首先，最直接的感受就是我们可以“轻装上阵”地开发复杂系统了。以前动不动就需要十几个人而且团队里需要不少技术达人一起分工协作才能设计实现和运维的分布式系统，在采用 Kubernetes 解决方案之后，只需一个精悍的小团队就能轻松应对。在这个团队里，一名架构师专注于系统中“服务组件”的提炼，几名开发工程师专注于业务代码的开发，一名系统兼运维工程师负责 Kubernetes 的部署和运维，从此再也不用“996”了，这并不是因为我们少做了什么，而是因为 Kubernetes 已经帮我们做了很多。

其次，使用 Kubernetes 就是在全面拥抱微服务架构。微服务架构的核心是将一个巨大的单体应用分解为很多小的互相连接的微服务，一个微服务背后可能有多个实例副本在支撑，副本的数量可能会随着系统的负荷变化而进行调整，内嵌的负载均衡器在这里发挥了重要作用。微服务架构使得每个服务都可以由专门的开发团队来开发，开发者可以自由选择开发技术，这对于大规模团队来说很有价值，另外每个微服务独立开发、升级、扩展，因此系统具备很高的稳定性和快速迭代进化能力。谷歌、亚马逊、eBay、Netflix 等众多大型互联网公司都采用了微服务架构，此次谷歌更是将微服务架构的基础设施直接打包到 Kubernetes 解决方案中，让我们有机会直接应用微服务架构解决复杂业务系统的架构问题。

然后，我们的系统可以随时随地整体“搬迁”到公有云上。Kubernetes 最初的目标就是运

行在谷歌自家的公有云 GCE 中，未来会支持更多的公有云及基于 OpenStack 的私有云。同时，在 Kubernetes 的架构方案中，底层网络的细节完全被屏蔽，基于服务的 Cluster IP 甚至都无须我们改变运行期的配置文件，就能将系统从物理机环境中无缝迁移到公有云中，或者在服务高峰期将部分服务对应的 Pod 副本放入公有云中以提升系统的吞吐量，不仅节省了公司的硬件投入，还大大改善了客户体验。我们所熟知的铁道部的 12306 购票系统，在春节高峰期就租用了阿里云进行分流。

最后，Kubernetes 系统架构具备了超强的横向扩容能力。对于互联网公司来说，用户规模就等价于资产，谁拥有更多的用户，谁就能在竞争中胜出，因此超强的横向扩容能力是互联网业务系统的关键指标之一。不用修改代码，一个 Kubernetes 集群即可从只包含几个 Node 的小集群平滑扩展到拥有上百个 Node 的大规模集群，我们利用 Kubernetes 提供的工具，甚至可以在线完成集群扩容。只要我们的微服务设计得好，结合硬件或者公有云资源的线性增加，系统就能够承受大量用户并发访问所带来的巨大压力。

### 1.3 从一个简单的例子开始

考虑到本书第 1 版中的 PHP+Redis 留言板的 Hello World 例子对于绝大多数刚接触 Kubernetes 的人来说比较复杂，难以顺利上手和实践，所以我们在此将这个例子替换成一个简单得多的 Java Web 应用，可以让新手快速上手和实践。

此 Java Web 应用的结构比较简单，是一个运行在 Tomcat 里的 Web App，如图 1.1 所示，JSP 页面通过 JDBC 直接访问 MySQL 数据库并展示数据。为了演示和简化的目的，只要程序正确连接到了数据库上，它就会自动完成对应的 Table 的创建与初始化数据的准备工作。所以，当我们通过浏览器访问此应用时，就会显示一个表格的页面，数据则来自数据库。

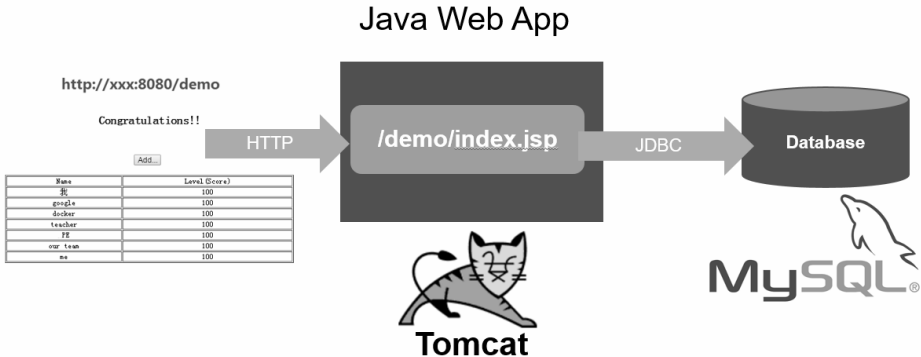


图 1.1 Java Web 应用的结构

此应用需要启动两个容器：Web App 容器和 MySQL 容器，并且 Web App 容器需要访问 MySQL 容器。在 Docker 时代，假设我们在一个宿主机上启动了这两个容器，则我们需要把 MySQL 容器的 IP 地址通过环境变量的方式注入 Web App 容器里；同时，需要将 Web App 容器的 8080 端口映射到宿主机的 8080 端口，以便能在外部访问。在本章的这个例子里，我们看看在 Kubernetes 时代是如何完成这个目标的。

### 1.3.1 环境准备

首先，我们开始准备 Kubernetes 的安装和相关镜像下载，本书建议采用 VirtualBox 或者 VMware Workstation 在本机虚拟一个 64 位的 CentOS 7 虚拟机作为学习环境，虚拟机采用 NAT 的网络模式以便能够连接外网，然后按照以下步骤快速安装 Kubernetes。

(1) 关闭 CentOS 自带的防火墙服务：

```
# systemctl disable firewalld
# systemctl stop firewalld
```

(2) 安装 etcd 和 Kubernetes 软件（会自动安装 Docker 软件）：

```
# yum install -y etcd kubernetes
```

(3) 按顺序启动所有的服务：

```
# systemctl start etcd
# systemctl start docker
# systemctl start kube-apiserver
# systemctl start kube-controller-manager
# systemctl start kube-scheduler
# systemctl start kubelet
# systemctl start kube-proxy
```

至此，一个单机版的 Kubernetes 集群环境就安装启动完成了。

接下来，我们可以在这个单机版的 Kubernetes 集群中上手练习了。

注：本书示例中的 Docker 镜像下载地址为 <https://hub.docker.com/u/kubeguide/>。

### 1.3.2 启动 MySQL 服务

首先为 MySQL 服务创建一个 RC 定义文件：mysql-rc.yaml，下面给出了该文件的完整内容和解释：

```
apiVersion: v1
kind: ReplicationController          # 副本控制器 RC
metadata:
```



```

name: mysql                                # RC 的名称，全局唯一
spec:
  replicas: 1                              # Pod 副本期待数量
  selector:
    app: mysql                             # 符合目标的 Pod 拥有此标签
  template:                                # 根据此模板创建 Pod 的副本（实例）
    metadata:
      labels:
        app: mysql                         # Pod 副本拥有的标签，对应 RC 的 Selector
    spec:
      containers:                          # Pod 内容器的定义部分
      - name: mysql                        # 容器的名称
        image: mysql                       # 容器对应的 Docker Image
        ports:
        - containerPort: 3306              # 容器应用监听的端口号
        env:                               # 注入容器内的环境变量
        - name: MYSQL_ROOT_PASSWORD
          value: "123456"

```

yaml 定义文件中的 `kind` 属性，用来表明此资源对象的类型，比如这里的值为“`ReplicationController`”，表示这是一个 RC；spec 一节中是 RC 的相关属性定义，比如 `spec.selector` 是 RC 的 Pod 标签（Label）选择器，即监控和管理拥有这些标签的 Pod 实例，确保当前集群上始终有且仅有 `replicas` 个 Pod 实例在运行，这里我们设置 `replicas=1` 表示只能运行一个 MySQL Pod 实例。当集群中运行的 Pod 数量小于 `replicas` 时，RC 会根据 `spec.template` 一节中定义的 Pod 模板来生成一个新的 Pod 实例，`spec.template.metadata.labels` 指定了该 Pod 的标签，需要特别注意的是：这里的 `labels` 必须匹配之前的 `spec.selector`，否则此 RC 每次创建了一个无法匹配 Label 的 Pod，就会不停地尝试创建新的 Pod，最终陷入“为他人作嫁衣裳”的悲惨世界中，永无翻身之时。

创建好 `mysql-rc.yaml` 文件以后，为了将它发布到 Kubernetes 集群中，我们在 Master 节点执行命令：

```

# kubectl create -f mysql-rc.yaml
replicationcontroller "mysql" created

```

接下来，我们用 `kubectl` 命令查看刚刚创建的 RC：

```

# kubectl get rc
NAME          DESIRED   CURRENT   AGE
mysql         1         1         1m

```

查看 Pod 的创建情况时，可以运行下面的命令：

```

# kubectl get pods
NAME           READY   STATUS    RESTARTS   AGE
mysql-c95jc    1/1     Running   0           2m

```

我们看到一个名为 `mysql-xxxxx` 的 Pod 实例，这是 Kubernetes 根据 `mysql` 这个 RC 的定义

自动创建的 Pod。由于 Pod 的调度和创建需要花费一定的时间，比如需要一定的时间来确定调度到哪个节点上，以及下载 Pod 里容器的镜像需要一段时间，所以一开始我们看到 Pod 的状态将显示为 Pending。当 Pod 成功创建完成以后，状态最终会被更新为 Running。

我们通过 `docker ps` 指令查看正在运行的容器，发现提供 MySQL 服务的 Pod 容器已经创建并正常运行了，此外，你会发现 MySQL Pod 对应的容器还多创建了一个来自谷歌的 `pause` 容器，这就是 Pod 的“根容器”，详见 1.4.3 节的说明。

```
# docker ps | grep mysql
72ca992535b4 mysql
"docker-entrypoint.sh" 12 minutes ago Up 12 minutes
k8s_mysql.86dc506e_mysql-c95jc_default_511d6705-5051-11e6-a9d8-000c29ed42c1_9f89d0b4
76c1790aad27 gcr.io/google_containers/pause-amd64:3.0
"/pause" 12 minutes ago Up 12 minutes
k8s_POD.16b20365_mysql-c95jc_default_511d6705-5051-11e6-a9d8-000c29ed42c1_28520aba
```

最后，我们创建一个与之关联的 Kubernetes Service——MySQL 的定义文件（文件名为 `mysql-svc.yaml`），完整的内容和解释如下：

```
apiVersion: v1
kind: Service # 表明是 Kubernetes Service
metadata:
  name: mysql # Service 的全局唯一名称
spec:
  ports:
    - port: 3306 # Service 提供服务的端口号
  selector: # Service 对应的 Pod 拥有这里定义的标签
    app: mysql
```

其中，`metadata.name` 是 Service 的服务名（ServiceName）；`port` 属性则定义了 Service 的虚端口；`spec.selector` 确定了哪些 Pod 副本（实例）对应到本服务。类似地，我们通过 `kubectl create` 命令创建 Service 对象。

运行 `kubectl` 命令，创建 service：

```
# kubectl create -f mysql-svc.yaml
service "mysql" created
```

运行 `kubectl` 命令，可以查看到刚刚创建的 service：

```
# kubectl get svc
NAME          CLUSTER-IP      EXTERNAL-IP      PORT(S)          AGE
mysql         169.169.253.143 <none>           3306/TCP         48s
```

注意到 MySQL 服务被分配了一个值为 169.169.253.143 的 Cluster IP 地址，这是一个虚地址，随后，Kubernetes 集群中其他新创建的 Pod 就可以通过 Service 的 Cluster IP+端口号 3306 来连接和访问它了。

在通常情况下，Cluster IP 是在 Service 创建后由 Kubernetes 系统自动分配的，其他 Pod 无法预先知道某个 Service 的 Cluster IP 地址，因此需要一个服务发现机制来找到这个服务。为此，最初时，Kubernetes 巧妙地使用了 Linux 环境变量（Environment Variable）来解决这个问题，后面会详细说明其机制。现在我们只需知道，根据 Service 的唯一名字，容器可以从环境变量中获取到 Service 对应的 Cluster IP 地址和端口，从而发起 TCP/IP 连接请求了。

### 1.3.3 启动 Tomcat 应用

上面我们定义和启动了 MySQL 服务，接下来我们采用同样的步骤，完成 Tomcat 应用的启动过程。首先，创建对应的 RC 文件 myweb-rc.yaml，内容如下：

```
apiVersion: v1
kind: ReplicationController
metadata:
  name: myweb
spec:
  replicas: 2
  selector:
    app: myweb
  template:
    metadata:
      labels:
        app: myweb
    spec:
      containers:
        - name: myweb
          image: kubeguide/tomcat-app:v1
          ports:
            - containerPort: 8080
```

注意：Tomcat 容器内，应用将使用环境变量 MYSQL\_SERVICE\_HOST 的值连接 MySQL 服务。更安全可靠的用法是使用服务的名称“mysql”，详见本章 Service 的概念和 2.4 节的说明。运行下面的命令，完成 RC 的创建和验证工作：

```
#kubectl create -f myweb-rc.yaml
replicationcontroller "myweb" created

# kubectl get pods
NAME          READY   STATUS    RESTARTS   AGE
mysql-c95jc   1/1     Running   0          2h
myweb-g9pmm   1/1     Running   0          3s
```

最后，创建对应的 Service。以下是完整的 yaml 定义文件（myweb-svc.yaml）：

```
apiVersion: v1
```

```
kind: Service
metadata:
  name: myweb
spec:
  type: NodePort
  ports:
    - port: 8080
      nodePort: 30001
  selector:
    app: myweb
```

`type=NodePort` 和 `nodePort=30001` 的两个属性，表明此 Service 开启了 NodePort 方式的外网访问模式，在 Kubernetes 集群之外，比如在本机的浏览器里，可以通过 30001 这个端口访问 myweb（对应到 8080 的虚端口上）。

运行 `kubectl create` 命令进行创建：

```
# kubectl create -f myweb-svc.yaml
You have exposed your service on an external port on all nodes in your
cluster. If you want to expose this service to the external internet, you may
need to set up firewall rules for the service port(s) (tcp:30001) to serve traffic.
See http://releases.k8s.io/release-1.3/docs/user-guide/services-firewalls.md
for more details.
service "myweb" created
```

我们看到上面有提示信息，意思是需要把 30001 这个端口在防火墙上打开，以便外部的访问能穿过防火墙。

运行 `kubectl` 命令，查看创建的 Service：

```
# kubectl get services
```

NAME	CLUSTER-IP	EXTERNAL-IP	PORT(S)	AGE
mysql	169.169.253.143	<none>	3306/TCP	2m
myweb	169.169.149.215	<nodes>	8080/TCP	1m
kubernetes	169.169.0.1	<none>	443/TCP	10m

至此，我们的第 1 个 Kubernetes 例子搭建完成了，在下一节中我们验证结果。

### 1.3.4 通过浏览器访问网页

经过上面的几个步骤，我们终于成功实现了 Kubernetes 上第 1 个例子的部署搭建工作。现在一起来见证成果吧，在你的笔记本上打开浏览器，输入 `http://虚拟机 IP:30001/demo/`。

比如虚拟机 IP 为 192.168.18.131（可以通过 `#ip a` 命令进行查询），在浏览器里输入地址 `http://192.168.18.131:30001/demo/` 后，看到了如图 1.2 所示的网页界面，那么恭喜你，之前的努力没有白费，顺利闯关成功！

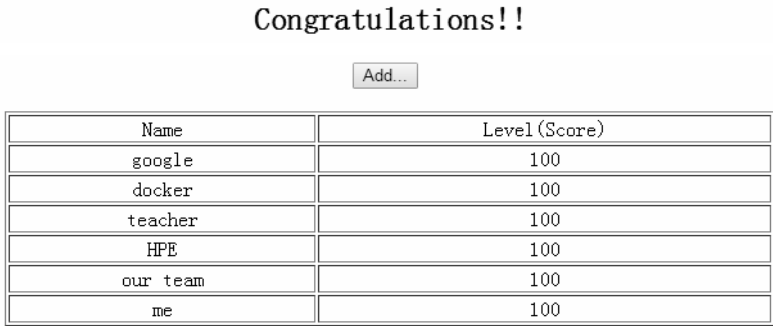


图 1.2 通过浏览器访问 Tomcat 应用

如果看不到这个网页，那么可能有几个原因：比如防火墙的问题，无法访问 30001 端口，或者因为你是通过代理上网的，浏览器错把虚拟机的 IP 地址当成远程地址了。可以在虚拟机上直接运行 `curl 192.168.18.131:30001` 来验证此端口是否能被访问，如果还是不能访问，那么这肯定不是机器的问题……

接下来可以尝试单击“Add...”按钮添加一条记录并提交，如图 1.3 所示，提交以后，数据就被写入 MySQL 数据库中了。

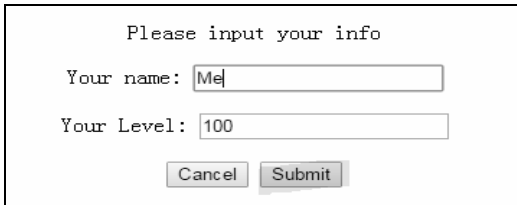


图 1.3 在留言板网页添加新的留言

至此，我们终于完成了 Kubernetes 上的 Tomcat 例子，这个例子并不是很复杂。我们也看到，相对于传统的分布式应用的部署方式，在 Kubernetes 之上我们仅仅通过一些很容易理解的配置文件和相关的简单命令就完成了对整个集群的部署，这让我们惊诧于 Kubernetes 的创新和强大。

下一节，我们将开始对 Kubernetes 中的基本概念和术语进行全面学习，在这之前，读者可以继续研究下这个例子里的一些拓展内容，如下所述。

- ◎ 研究 RC、Service 等配置文件的格式。
- ◎ 熟悉 kubectl 的子命令。
- ◎ 手工停止某个 Service 对应的容器进程，然后观察有什么现象发生。
- ◎ 修改 RC 文件，改变副本数量，重新发布，观察结果。

## 1.4 Kubernetes 基本概念和术语

Kubernetes 中的大部分概念如 Node、Pod、Replication Controller、Service 等都可以看作一种“资源对象”，几乎所有的资源对象都可以通过 Kubernetes 提供的 `kubectl` 工具（或者 API 编程调用）执行增、删、改、查等操作并将其保存在 `etcd` 中持久化存储。从这个角度来看，Kubernetes 其实是一个高度自动化的资源控制系统，它通过跟踪对比 `etcd` 库里保存的“资源期望状态”与当前环境中的“实际资源状态”的差异来实现自动控制和自动纠错的高级功能。

在介绍资源对象之前，我们先了解一下 Kubernetes 集群的两种管理角色：Master 和 Node。

### 1.4.1 Master

Kubernetes 里的 Master 指的是集群控制节点，每个 Kubernetes 集群里需要有一个 Master 节点来负责整个集群的管理和控制，基本上 Kubernetes 的所有控制命令都发给它，它来负责具体的执行过程，我们后面执行的所有命令基本都是在 Master 节点上运行的。Master 节点通常会占据一个独立的服务器（高可用部署建议用 3 台服务器），其主要原因是它太重要了，是整个集群的“首脑”，如果宕机或者不可用，那么对集群容器应用的管理都将失效。

Master 节点上运行着以下一组关键进程。

- ◎ Kubernetes API Server (`kube-apiserver`)：提供了 HTTP Rest 接口的关键服务进程，是 Kubernetes 里所有资源的增、删、改、查等操作的唯一入口，也是集群控制的入口进程。
- ◎ Kubernetes Controller Manager (`kube-controller-manager`)：Kubernetes 里所有资源对象的自动化控制中心，可以理解为资源对象的“大总管”。
- ◎ Kubernetes Scheduler (`kube-scheduler`)：负责资源调度（Pod 调度）的进程，相当于公交公司的“调度室”。

另外，在 Master 节点上还需要启动一个 `etcd` 服务，因为 Kubernetes 里的所有资源对象的数据全部是保存在 `etcd` 中的。

### 1.4.2 Node

除了 Master，Kubernetes 集群中的其他机器被称为 Node 节点，在较早的版本中也被称为 Minion。与 Master 一样，Node 节点可以是一台物理主机，也可以是一台虚拟机。Node 节点才是 Kubernetes 集群中的工作负载节点，每个 Node 都会被 Master 分配一些工作负载（Docker 容

器), 当某个 Node 宕机时, 其上的工作负载会被 Master 自动转移到其他节点上去。

每个 Node 节点上都运行着以下一组关键进程。

- ◎ **kubelet**: 负责 Pod 对应的容器的创建、启停等任务, 同时与 Master 节点密切协作, 实现集群管理的基本功能。
- ◎ **kube-proxy**: 实现 Kubernetes Service 的通信与负载均衡机制的重要组件。
- ◎ **Docker Engine (docker)**: Docker 引擎, 负责本机的容器创建和管理工作。

Node 节点可以在运行期间动态增加到 Kubernetes 集群中, 前提是这个节点上已经正确安装、配置和启动了上述关键进程, 在默认情况下 kubelet 会向 Master 注册自己, 这也是 Kubernetes 推荐的 Node 管理方式。一旦 Node 被纳入集群管理范围, kubelet 进程就会定时向 Master 节点汇报自身的情报, 例如操作系统、Docker 版本、机器的 CPU 和内存情况, 以及当前有哪些 Pod 在运行等, 这样 Master 可以获知每个 Node 的资源使用情况, 并实现高效均衡的资源调度策略。而某个 Node 超过指定时间不上报信息时, 会被 Master 判定为“失联”, Node 的状态被标记为不可用 (Not Ready), 随后 Master 会触发“工作负载大转移”的自动流程。

我们可以执行下述命令查看集群中有多少个 Node:

```
# kubectl get nodes
NAME                STATUS      AGE
kubernetes-minion1  Ready      2d
```

然后, 通过 `kubectl describe node <node_name>` 来查看某个 Node 的详细信息:

```
$ kubectl describe node kubernetes-minion1

Name:          k8s-node-1
Labels:        beta.kubernetes.io/arch=amd64
               beta.kubernetes.io/os=linux
               kubernetes.io/hostname=k8s-node-1
Taints:        <none>
CreationTimestamp:  Wed, 06 Jul 2016 11:46:41 +0800
Phase:
Conditions:
  Type             Status LastHeartbeatTime             Wed      xxxx
  OutOfDisk        False Sat, 09 Jul 2016 08:17:39 +0800  Wed      ....
  MemoryPressure   False Sat, 09 Jul 2016 08:17:39 +0800  Wed      .....
  Ready            True  Sat, 09 Jul 2016 08:17:39 +0800  Wed      .....
Addresses:       192.168.18.131,192.168.18.131
Capacity:
  alpha.kubernetes.io/nvidia-gpu: 0
  cpu:                             4
  memory:                          1868692Ki
  pods:                             110
```

```
Allocatable:
  alpha.kubernetes.io/nvidia-gpu:    0
  cpu:                                4
  memory:                             1868692Ki
  pods:                               110
System Info:
  Machine ID:                         6e4e2af2afeb42b9aac47d866aa56ca0
  System UUID:                       564D63D3-9664-3393-A3DC-9CD424ED42C1
  Boot ID:                           b0c34f9f-76ab-478e-9771-bd4fe6e98880
  Kernel Version:                    3.10.0-327.22.2.el7.x86_64
  OS Image:                          CentOS Linux 7 (Core)
  Operating System:                  linux
  Architecture:                      amd64
  Container Runtime Version: docker://1.11.2
  Kubelet Version:                   v1.3.0
  Kube-Proxy Version:                v1.3.0
  ExternalID:                        k8s-node-1
  Non-terminated Pods:               (1 in total)
    Namespace      Name      CPU Requests  CPU Limits Memory xxx
    -----
    kube-system    kube-dns-v11-wxdhf      310m (7%)   310m (7%)  170Mi (9%)
Allocated resources:
  (Total limits may be over 100 percent, i.e., overcommitted. More info:
  CPU Requests      CPU Limits Memory Requests      Memory Limits
  -----
    310m (7%)      310m (7%)   170Mi (9%)      170Mi (9%)
No events.
```

上述命令展示了 Node 的如下关键信息。

- ◎ Node 基本信息：名称、标签、创建时间等。
- ◎ Node 当前的运行状态，Node 启动以后会做一系列的自检工作，比如磁盘是否满了，如果满了就标注 `OutOfDisk=True`，否则继续检查内存是否不足（如果内存不足，就标注 `MemoryPressure=True`），最后一切正常，就设置为 Ready 状态（`Ready=True`），该状态表示 Node 处于健康状态，Master 将可以在其上调度新的任务了（如启动 Pod）。
- ◎ Node 的主机地址与主机名。
- ◎ Node 上的资源总量：描述 Node 可用的系统资源，包括 CPU、内存数量、最大可调度 Pod 数量等，注意到目前 Kubernetes 已经实验性地支持 GPU 资源分配了（`alpha.kubernetes.io/nvidia-gpu=0`）。
- ◎ Node 可分配资源量：描述 Node 当前可用于分配的资源量。
- ◎ 主机系统信息：包括主机的唯一标识 UUID、Linux kernel 版本号、操作系统类型与版本、Kubernetes 版本号、kubelet 与 kube-proxy 的版本号等。



- ◎ 当前正在运行的 Pod 列表概要信息。
- ◎ 已分配的资源使用概要信息，例如资源申请的最低、最大允许使用量占系统总量的百分比。
- ◎ Node 相关的 Event 信息。

### 1.4.3 Pod

Pod 是 Kubernetes 的最重要也最基本的概念，如图 1.4 所示是 Pod 的组成示意图，我们看到每个 Pod 都有一个特殊的被称为“根容器”的 Pause 容器。Pause 容器对应的镜像属于 Kubernetes 平台的一部分，除了 Pause 容器，每个 Pod 还包含一个或多个紧密相关的用户业务容器。

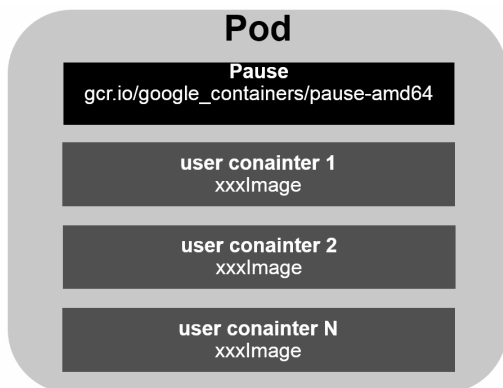


图 1.4 Pod 的组成

为什么 Kubernetes 会设计出一个全新的 Pod 的概念并且 Pod 有这样特殊的组成结构？

原因之一：在一组容器作为一个单元的情况下，我们难以对“整体”简单地进行判断及有效地进行行动。比如，一个容器死亡了，此时算是整体死亡么？是  $N/M$  的死亡率么？引入业务无关并且不易死亡的 Pause 容器作为 Pod 的根容器，以它的状态代表整个容器组的状态，就简单、巧妙地解决了这个难题。

原因之二：Pod 里的多个业务容器共享 Pause 容器的 IP，共享 Pause 容器挂接的 Volume，这样既简化了密切关联的业务容器之间的通信问题，也很好解决了它们之间的文件共享问题。

Kubernetes 为每个 Pod 都分配了唯一的 IP 地址，称之为 Pod IP，一个 Pod 里的多个容器共享 Pod IP 地址。Kubernetes 要求底层网络支持集群内任意两个 Pod 之间的 TCP/IP 直接通信，这通常采用虚拟二层网络技术来实现，例如 Flannel、Open vSwitch 等，因此我们需要牢记一点：在 Kubernetes 里，一个 Pod 里的容器与另外主机上的 Pod 容器能够直接通信。

Pod 其实有两种类型：普通的 Pod 及静态 Pod（Static Pod），后者比较特殊，它并不存放在

Kubernetes 的 etcd 存储里，而是存放在某个具体的 Node 上的一个具体文件中，并且只在此 Node 上启动运行。而普通的 Pod 一旦被创建，就会被放入到 etcd 中存储，随后会被 Kubernetes Master 调度到某个具体的 Node 上并进行绑定（Binding），随后该 Pod 被对应的 Node 上的 kubelet 进程实例化成一组相关的 Docker 容器并启动起来。在默认情况下，当 Pod 里的某个容器停止时，Kubernetes 会自动检测到这个问题并且重新启动这个 Pod（重启 Pod 里的所有容器），如果 Pod 所在的 Node 宕机，则会将这个 Node 上的所有 Pod 重新调度到其他节点上。Pod、容器与 Node 的关系如图 1.5 所示。

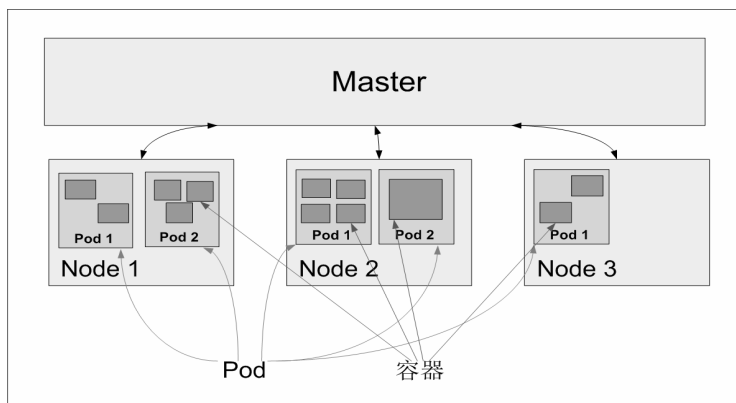


图 1.5 Pod、容器与 Node 的关系

Kubernetes 里的所有资源对象都可以采用 yaml 或者 JSON 格式的文件来定义或描述，下面是我们在之前 Hello World 例子里用到的 myweb 这个 Pod 的资源定义文件：

```
apiVersion: v1
kind: Pod
metadata:
  name: myweb
  labels:
    name: myweb
spec:
  containers:
  - name: myweb
    image: kubeguide/tomcat-app:v1
    ports:
    - containerPort: 8080
    env:
    - name: MYSQL_SERVICE_HOST
      value: 'mysql'
    - name: MYSQL_SERVICE_PORT
      value: '3306'
```

Kind 为 Pod 表明这是一个 Pod 的定义，metadata 里的 name 属性为 Pod 的名字，metadata 里还能定义资源对象的标签 (Label)，这里声明 myweb 拥有一个 name=myweb 的标签 (Label)。Pod 里所包含的容器组的定义则在 spec 一节中声明，这里定义了一个名字为 myweb、对应镜像为 kubeguide/tomcat-app:v1 的容器，该容器注入了名为 MYSQL\_SERVICE\_HOST='mysql' 和 MYSQL\_SERVICE\_PORT='3306' 的环境变量 (env 关键字)，并且在 8080 端口 (containerPort) 上启动容器进程。Pod 的 IP 加上这里的容器端口 (containerPort)，就组成了一个新的概念——Endpoint，它代表着此 Pod 里的一个服务进程的对外通信地址。一个 Pod 也存在着具有多个 Endpoint 的情况，比如当我们把 Tomcat 定义为一个 Pod 时，可以对外暴露管理端口与服务端口这两个 Endpoint。

我们所熟悉的 Docker Volume 在 Kubernetes 里也有对应的概念——Pod Volume，后者有一些扩展，比如可以用分布式文件系统 GlusterFS 实现后端存储功能；Pod Volume 是定义在 Pod 之上，然后被各个容器挂载到自己的文件系统上的。

这里顺便提一下 Kubernetes 的 Event 概念，Event 是一个事件的记录，记录了事件的最早产生时间、最后重现时间、重复次数、发起者、类型，以及导致此事件的原因等众多信息。Event 通常会关联到某个具体的资源对象上，是排查故障的重要参考信息，之前我们看到 Node 的描述信息包括了 Event，而 Pod 同样有 Event 记录，当我们发现某个 Pod 迟迟无法创建时，可以用 `kubectl describe pod xxxx` 来查看它的描述信息，用来定位问题的原因，比如下面这个 Event 记录信息表明 Pod 里的一个容器被探针检测为失败一次：

```
Events:
  FirstSeen    LastSeen  Count  From              SubobjectPath  Type      Reason
  Message
  -----
10h          12m       32    {kubelet k8s-node-1} spec.containers{kube2sky}
Warning      Unhealthy  Liveness probe failed: Get http://172.17.1.2:8080/healthz:
net/http: request canceled (Client.Timeout exceeded while awaiting headers)
```

每个 Pod 都可以对其能使用的服务器上的计算资源设置限额，当前可以设置限额的计算资源有 CPU 与 Memory 两种，其中 CPU 的资源单位为 CPU (Core) 的数量，是一个绝对值而非相对值。

一个 CPU 的配额对于绝大多数容器来说是相当大的一个资源配额了，所以，在 Kubernetes 里，通常以千分之一的 CPU 配额为最小单位，用 m 来表示。通常一个容器的 CPU 配额被定义为 100~300m，即占用 0.1~0.3 个 CPU。由于 CPU 配额是一个绝对值，所以无论在拥有一个 Core 的机器上，还是在拥有 48 个 Core 的机器上，100m 这个配额所代表的 CPU 的使用量都是一样的。与 CPU 配额类似，Memory 配额也是一个绝对值，它的单位是内存字节数。

在 Kubernetes 里，一个计算资源进行配额限定需要设定以下两个参数。

- ◎ **Requests**: 该资源的最小申请量，系统必须满足要求。
- ◎ **Limits**: 该资源最大允许使用的量，不能被突破，当容器试图使用超过这个量的资源时，可能会被 Kubernetes Kill 并重启。

通常我们会把 Request 设置为一个比较小的数值，符合容器平时的工作负载情况下的资源需求，而把 Limit 设置为峰值负载情况下资源占用的最大量。比如下面这段定义，表明 MySQL 容器申请最少 0.25 个 CPU 及 64MiB 内存，在运行过程中 MySQL 容器所能使用的资源配额为 0.5 个 CPU 及 128MiB 内存：

```
spec:
  containers:
  - name: db
    image: mysql
    resources:
      requests:
        memory: "64Mi"
        cpu: "250m"
      limits:
        memory: "128Mi"
        cpu: "500m"
```

本节最后，笔者给出 Pod 及 Pod 周边对象的示意图作为总结，如图 1.6 所示，后面部分还会涉及这张图里的对象和概念，以进一步加强理解。

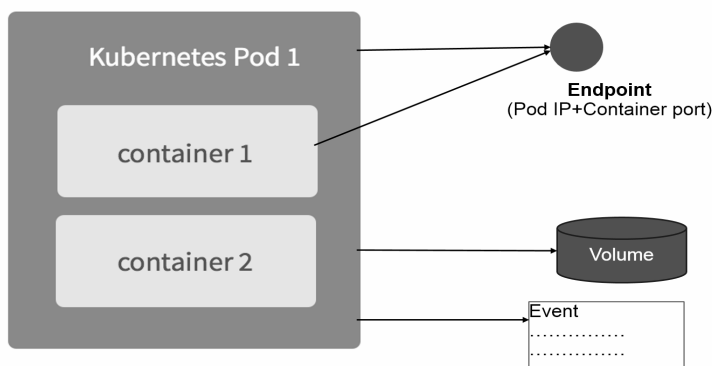


图 1.6 Pod 及周边对象

#### 1.4.4 Label（标签）

Label 是 Kubernetes 系统中另外一个核心概念。一个 Label 是一个 key=value 的键值对，

其中 key 与 value 由用户自己指定。Label 可以附加到各种资源对象上，例如 Node、Pod、Service、RC 等，一个资源对象可以定义任意数量的 Label，同一个 Label 也可以被添加到任意数量的资源对象上去，Label 通常在资源对象定义时确定，也可以在对象创建后动态添加或者删除。

我们可以通过给指定的资源对象捆绑一个或多个不同的 Label 来实现多维度的资源分组管理功能，以便于灵活、方便地进行资源分配、调度、配置、部署等管理工作。例如：部署不同版本的应用到不同的环境中；或者监控和分析应用（日志记录、监控、告警）等。一些常用的 Label 示例如下。

- ◎ 版本标签: "release": "stable", "release": "canary"...
- ◎ 环境标签: "environment": "dev", "environment": "qa", "environment": "production"
- ◎ 架构标签: "tier": "frontend", "tier": "backend", "tier": "middleware"
- ◎ 分区标签: "partition": "customerA", "partition": "customerB"...
- ◎ 质量管控标签: "track": "daily", "track": "weekly"

Label 相当于我们熟悉的“标签”，给某个资源对象定义一个 Label，就相当于给它打了一个标签，随后可以通过 Label Selector（标签选择器）查询和筛选拥有某些 Label 的资源对象，Kubernetes 通过这种方式实现了类似 SQL 的简单又通用的对象查询机制。

Label Selector 可以被类比为 SQL 语句中的 where 查询条件，例如，name=redis-slave 这个 Label Selector 作用于 Pod 时，可以被类比为 select \* from pod where pod's name = 'redis-slave' 这样的语句。当前有两种 Label Selector 的表达式：基于等式的（Equality-based）和基于集合的（Set-based），前者采用“等式类”的表达式匹配标签，下面是一些具体的例子。

- ◎ name = redis-slave: 匹配所有具有标签 name=redis-slave 的资源对象。
- ◎ env != production: 匹配所有不具有标签 env=production 的资源对象，比如 env=test 就是满足此条件的标签之一。

而后者则使用集合操作的表达式匹配标签，下面是一些具体的例子。

- ◎ name in (redis-master, redis-slave): 匹配所有具有标签 name=redis-master 或者 name=redis-slave 的资源对象。
- ◎ name not in (php-frontend): 匹配所有不具有标签 name=php-frontend 的资源对象。

可以通过多个 Label Selector 表达式的组合实现复杂的条件选择，多个表达式之间用“，”进行分隔即可，几个条件之间是“AND”的关系，即同时满足多个条件，比如下面的例子：

```
name=redis-slave,env!=production
name notin (php-frontend),env!=production
```

以 myweb Pod 为例，Label 定义在其 metadata 中：

```
apiVersion: v1
kind: Pod
metadata:
  name: myweb
  labels:
    app: myweb
```

管理对象 RC 和 Service 在 spec 中定义 Selector 与 Pod 进行关联：

```
apiVersion: v1
kind: ReplicationController
metadata:
  name: myweb
spec:
  replicas: 1
  selector:
    app: myweb
  template:
    ...略...
```

```
apiVersion: v1
kind: Service
metadata:
  name: myweb
spec:
  selector:
    app: myweb
  ports:
    - port: 8080
```

新出现的管理对象如 Deployment、ReplicaSet、DaemonSet 和 Job 则可以在 Selector 中使用基于集合的筛选条件定义，例如：

```
selector:
  matchLabels:
    app: myweb
  matchExpressions:
    - {key: tier, operator: In, values: [frontend]}
    - {key: environment, operator: NotIn, values: [dev]}
```

matchLabels 用于定义一组 Label，与直接写在 Selector 中作用相同；matchExpressions 用于定义一组基于集合的筛选条件，可用的条件运算符包括：In、NotIn、Exists 和 DoesNotExist。

如果同时设置了 matchLabels 和 matchExpressions，则两组条件为“AND”关系，即所有条件需要同时满足才能完成 Selector 的筛选。

Label Selector 在 Kubernetes 中的重要使用场景有以下几处。

- ◎ kube-controller 进程通过资源对象 RC 上定义的 Label Selector 来筛选要监控的 Pod 副本的数量，从而实现 Pod 副本的数量始终符合预期设定的全自动控制流程。
- ◎ kube-proxy 进程通过 Service 的 Label Selector 来选择对应的 Pod，自动建立起每个 Service 到对应 Pod 的请求转发路由表，从而实现 Service 的智能负载均衡机制。
- ◎ 通过对某些 Node 定义特定的 Label，并且在 Pod 定义文件中使用 NodeSelector 这种标签调度策略，kube-scheduler 进程可以实现 Pod “定向调度”的特性。

在前面的留言板例子中，我们只使用了一个 name=XXX 的 Label Selector。让我们看一个更复杂的例子。假设为 Pod 定义了 3 个 Label: release、env 和 role，不同的 Pod 定义了不同的 Label 值，如图 1.7 所示，如果我们设置了“role=frontend”的 Label Selector，则会选取到 Node 1 和 Node 2 上的 Pod。

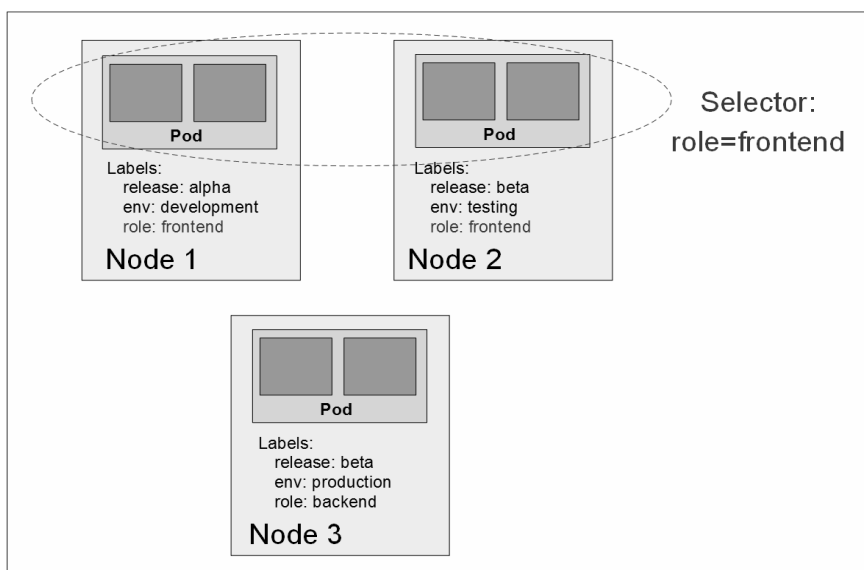


图 1.7 Label Selector 的作用范围 1

而设置“release=beta”的 Label Selector，则会选取到 Node 2 和 Node 3 上的 Pod，如图 1.8 所示。

总结：使用 Label 可以给对象创建多组标签，Label 和 Label Selector 共同构成了 Kubernetes 系统中最核心的应用模型，使得被管理对象能够被精细地分组管理，同时实现了整个集群的高可用性。

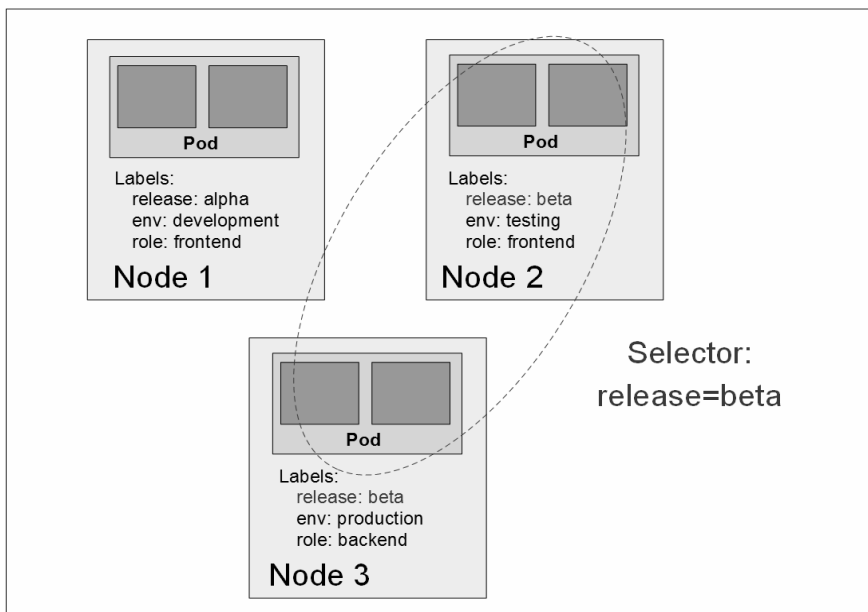


图 1.8 Label Selector 的作用范围 2

### 1.4.5 Replication Controller

上一节的例子中已经对 Replication Controller（简称 RC）的定义和作用做了一些说明，本节对 RC 的概念进行深入描述。

RC 是 Kubernetes 系统中的核心概念之一，简单来说，它其实是定义了一个期望的场景，即声明某种 Pod 的副本数量在任意时刻都符合某个预期值，所以 RC 的定义包括如下几个部分。

- ◎ Pod 期待的副本数（replicas）。
- ◎ 用于筛选目标 Pod 的 Label Selector。
- ◎ 当 Pod 的副本数量小于预期数量时，用于创建新 Pod 的 Pod 模板（template）。

下面是一个完整的 RC 定义的例子，即确保拥有 tier=frontend 标签的这个 Pod（运行 Tomcat 容器）在整个 Kubernetes 集群中始终只有一个副本：

```
apiVersion: v1
kind: ReplicationController
metadata:
  name: frontend
spec:
  replicas: 1
```



```

selector:
  tier: frontend
template:
  metadata:
    labels:
      app: app-demo
      tier: frontend
  spec:
    containers:
      - name: tomcat-demo
        image: tomcat
        imagePullPolicy: IfNotPresent
        env:
          - name: GET_HOSTS_FROM
            value: dns
        ports:
          - containerPort: 80

```

当我们定义了一个 RC 并提交到 Kubernetes 集群中以后, Master 节点上的 Controller Manager 组件就得到通知, 定期巡检系统中当前存活的目标 Pod, 并确保目标 Pod 实例的数量刚好等于此 RC 的期望值, 如果有过多的 Pod 副本在运行, 系统就会停掉一些 Pod, 否则系统就会再自动创建一些 Pod。可以说, 通过 RC, Kubernetes 实现了用户应用集群的高可用性, 并且大大减少了系统管理员在传统 IT 环境中需要完成的许多手工运维工作(如主机监控脚本、应用监控脚本、故障恢复脚本等)。

下面我们以 3 个 Node 节点的集群为例, 说明 Kubernetes 如何通过 RC 来实现 Pod 副本数量自动控制的机制。假如我们的 RC 里定义 redis-slave 这个 Pod 需要保持 3 个副本, 系统将可能在其中的两个 Node 上创建 Pod。图 1.9 描述了在两个 Node 上创建 redis-slave Pod 的情形。

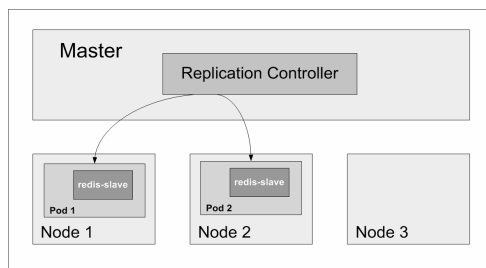


图 1.9 在两个 Node 上创建 redis-slave Pod

假设 Node 2 上的 Pod 2 意外终止, 根据 RC 定义的 replicas 数量 2, Kubernetes 将会自动创建并启动一个新的 Pod, 以保证整个集群中始终有两个 redis-slave Pod 在运行。

如图 1.10 所示, 系统可能选择 Node 3 或者 Node 1 来创建一个新的 Pod。

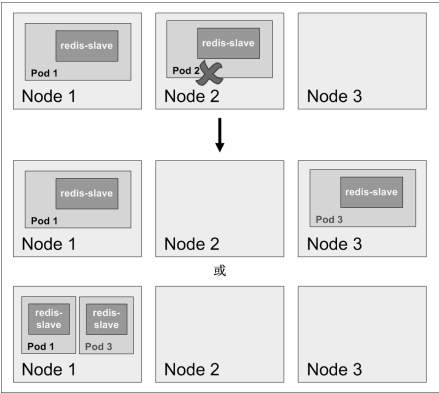


图 1.10 根据 RC 定义创建新的 Pod

此外，在运行时，我们可以通过修改 RC 的副本数量，来实现 Pod 的动态缩放（Scaling）功能，这可以通过执行 `kubectl scale` 命令来一键完成：

```
$ kubectl scale rc redis-slave --replicas=3
scaled
```

Scaling 的执行结果如图 1.11 所示。

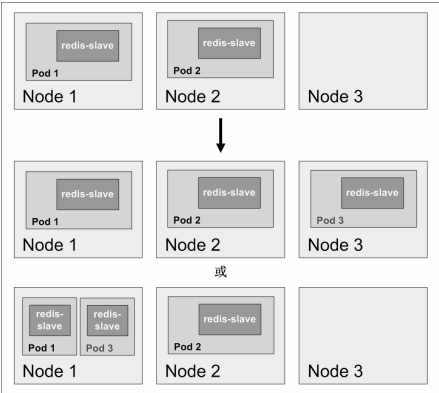


图 1.11 Scaling 的执行结果

需要注意的是，删除 RC 并不会影响通过该 RC 已创建好的 Pod。为了删除所有 Pod，可以设置 `replicas` 的值为 0，然后更新该 RC。另外，`kubectl` 提供了 `stop` 和 `delete` 命令来一次性删除 RC 和 RC 控制的全部 Pod。

当我们的应用升级时，通常会通过 Build 一个新的 Docker 镜像，并用新的镜像版本来替代旧的版本的方式达到目的。在系统升级的过程中，我们希望是平滑的方式，比如当前系统中 10 个对应的旧版本的 Pod，最佳的方式是旧版本的 Pod 每次停止一个，同时创建一个新版本的 Pod，

在整个升级过程中，此消彼长，而运行中的 Pod 数量始终是 10 个，几分钟以后，当所有的 Pod 都已经是新版本时，升级过程完成。通过 RC 的机制，Kubernetes 很容易就实现了这种高级实用的特性，被称为“滚动升级”（Rolling Update），具体的操作方法详见第 4 章。

由于 Replication Controller 与 Kubernetes 代码中的模块 Replication Controller 同名，同时这个词也无法准确表达它的本意，所以在 Kubernetes v1.2 时，它就升级成了另外一个新的概念——Replica Set，官方解释为“下一代的 RC”，它与 RC 当前存在的唯一区别是：Replica Sets 支持基于集合的 Label selector（Set-based selector），而 RC 只支持基于等式的 Label Selector（equality-based selector），这使得 Replica Set 的功能更强，下面是等价于之前 RC 例子的 Replica Set 的定义（省去了 Pod 模板部分的内容）：

```
apiVersion: extensions/v1beta1
kind: ReplicaSet
metadata:
  name: frontend
spec:
  selector:
    matchLabels:
      tier: frontend
    matchExpressions:
      - {key: tier, operator: In, values: [frontend]}
  template:
    .....
```

kubectl 命令行工具适用于 RC 的绝大部分命令都同样适用于 Replica Set。此外，当前我们很少单独使用 Replica Set，它主要被 Deployment 这个更高层的资源对象所使用，从而形成一整套 Pod 创建、删除、更新的编排机制。当我们使用 Deployment 时，无须关心它是如何创建和维护 Replica Set 的，这一切都是自动发生的。

Replica Set 与 Deployment 这两个重要资源对象逐步替换了之前的 RC 的作用，是 Kubernetes v1.3 里 Pod 自动扩容（伸缩）这个告警功能实现的基础，也将继续在 Kubernetes 未来的版本中发挥重要的作用。

最后我们总结一下关于 RC（Replica Set）的一些特性与作用。

- ◎ 在大多数情况下，我们通过定义一个 RC 实现 Pod 的创建过程及副本数量的自动控制。
- ◎ RC 里包括完整的 Pod 定义模板。
- ◎ RC 通过 Label Selector 机制实现对 Pod 副本的自动控制。
- ◎ 通过改变 RC 里的 Pod 副本数量，可以实现 Pod 的扩容或缩容功能。
- ◎ 通过改变 RC 里 Pod 模板中的镜像版本，可以实现 Pod 的滚动升级功能。

## 1.4.6 Deployment

Deployment 是 Kubernetes v1.2 引入的新概念，引入的目的是为了更好地解决 Pod 的编排问题。为此，Deployment 在内部使用了 Replica Set 来实现目的，无论从 Deployment 的作用与目的、它的 YAML 定义，还是从它的具体命令行操作来看，我们都可以把它看作 RC 的一次升级，两者的相似度超过 90%。

Deployment 相对于 RC 的一个最大升级是我们可以随时知道当前 Pod “部署”的进度。实际上由于一个 Pod 的创建、调度、绑定节点及在目标 Node 上启动对应的容器这一完整过程需要一定的时间，所以我们期待系统启动  $N$  个 Pod 副本的目标状态，实际上是一个连续变化的“部署过程”导致的最终状态。

Deployment 的典型使用场景有以下几个。

- ◎ 创建一个 Deployment 对象来生成对应的 Replica Set 并完成 Pod 副本的创建过程。
- ◎ 检查 Deployment 的状态来看部署动作是否完成（Pod 副本的数量是否达到预期的值）。
- ◎ 更新 Deployment 以创建新的 Pod（比如镜像升级）。
- ◎ 如果当前 Deployment 不稳定，则回滚到一个早先的 Deployment 版本。
- ◎ 暂停 Deployment 以便于一次性修改多个 PodTemplateSpec 的配置项，之后再恢复 Deployment，进行新的发布。
- ◎ 扩展 Deployment 以应对高负载。
- ◎ 查看 Deployment 的状态，以此作为发布是否成功的指标。
- ◎ 清理不再需要的旧版本 ReplicaSets。

Deployment 的定义与 Replica Set 的定义很类似，除了 API 声明与 Kind 类型等有所区别：

apiVersion: extensions/v1beta1	apiVersion: v1
kind: Deployment	kind: ReplicaSet
metadata:	metadata:
name: nginx-deployment	name: nginx-repset

下面我们通过运行一些例子来一起直观地感受这个新概念。首先创建一个名为 tomcat-deployment.yaml 的 Deployment 描述文件，内容如下：

```
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: frontend
spec:
  replicas: 1
  selector:
```

```

matchLabels:
  tier: frontend
matchExpressions:
  - {key: tier, operator: In, values: [frontend]}
template:
  metadata:
    labels:
      app: app-demo
      tier: frontend
  spec:
    containers:
      - name: tomcat-demo
        image: tomcat
        imagePullPolicy: IfNotPresent
        ports:
          - containerPort: 8080

```

运行下述命令创建 Deployment:

```

# kubectl create -f tomcat-deployment.yaml
deployment "tomcat-deploy" created

```

运行下述命令查看 Deployment 的信息:

```

# kubectl get deployments

```

NAME	DESIRED	CURRENT	UP-TO-DATE	AVAILABLE	AGE
tomcat-deploy	1	1	1	1	4m

对上述输出中涉及的数量解释如下。

- ◎ **DESIRED:** Pod 副本数量的期望值, 即 Deployment 里定义的 Replica。
- ◎ **CURRENT:** 当前 Replica 的值, 实际上是 Deployment 所创建的 Replica Set 里的 Replica 值, 这个值不断增加, 直到达到 DESIRED 为止, 表明整个部署过程完成。
- ◎ **UP-TO-DATE:** 最新版本的 Pod 的副本数量, 用于指示在滚动升级的过程中, 有多少个 Pod 副本已经成功升级。
- ◎ **AVAILABLE:** 当前集群中可用的 Pod 副本数量, 即集群中当前存活的 Pod 数量。

运行下述命令查看对应的 Replica Set, 我们看到它的命名与 Deployment 的名字有关系:

```

# kubectl get rs

```

NAME	DESIRED	CURRENT	AGE
tomcat-deploy-1640611518	1	1	1m

运行下述命令查看创建的 Pod, 我们发现 Pod 的命名以 Deployment 对应的 Replica Set 的名字为前缀, 这种命名很清晰地表明了一个 Replica Set 创建了哪些 Pod, 对于 Pod 滚动升级这种复杂的过程来说, 很容易排查错误:

```
# kubectl get pods
```

NAME	READY	STATUS	RESTARTS	AGE
tomcat-deploy-1640611518-zhrsc	1/1	Running	0	3m

运行 `kubectl describe deployments`，可以清楚地看到 Deployment 控制的 Pod 的水平扩展过程，详见第 2 章的说明，这里不再赘述。

Pod 的管理对象，除了 RC 和 Deployment，还包括 ReplicaSet、DaemonSet、StatefulSet、Job 等，分别用于不同的应用场景中，将在第 2 章中进行详细介绍。

### 1.4.7 Horizontal Pod Autoscaler

在前两节提到过，通过手工执行 `kubectl scale` 命令，我们可以实现 Pod 扩容或缩容。如果仅仅至此为止，显然不符合谷歌对 Kubernetes 的定位目标——自动化、智能化。在谷歌看来，分布式系统要能够根据当前负载的变化情况自动触发水平扩展或缩容的行为，因为这一过程可能是频繁发生的、不可预料的，所以手动控制的方式是不现实的。

因此，Kubernetes 的 v1.0 版本实现后，这帮大牛们就已经在默默研究 Pod 智能扩容的特性了，并在 Kubernetes v1.1 的版本中首次发布这一重量级新特性——Horizontal Pod Autoscaling（Pod 横向自动扩容，简称 HPA）。随后的 v1.2 版本中 HPA 被升级为稳定版本（apiVersion: autoscaling/v1），但同时仍然保留旧版本（apiVersion: extensions/v1beta1）。从 v1.6 版本开始，对根据应用自定义指标进行自动扩容和缩容的功能进行增强，API 版本为 autoscaling/v2alpha1，仍在不断演进过程中。

HPA 与之前的 RC、Deployment 一样，也属于一种 Kubernetes 资源对象。通过追踪分析 RC 控制的所有目标 Pod 的负载变化情况，来确定是否需要针对性地调整目标 Pod 的副本数，这是 HPA 的实现原理。当前，HPA 可以有以下两种方式作为 Pod 负载的度量指标。

- ◎ CPUUtilizationPercentage。
- ◎ 应用程序自定义的度量指标，比如服务在每秒内的相应的请求数（TPS 或 QPS）。

CPUUtilizationPercentage 是一个算术平均值，即目标 Pod 所有副本自身的 CPU 利用率的平均值。一个 Pod 自身的 CPU 利用率是该 Pod 当前 CPU 的使用量除以它的 Pod Request 的值，比如我们定义一个 Pod 的 Pod Request 为 0.4，而当前 Pod 的 CPU 使用量为 0.2，则它的 CPU 使用率为 50%，如此一来，我们就可以就算出来一个 RC 控制的所有 Pod 副本的 CPU 利用率的算术平均值了。如果某一时刻 CPUUtilizationPercentage 的值超过 80%，则意味着当前的 Pod 副本数很可能不足以支撑接下来更多的请求，需要进行动态扩容，而当请求高峰时段过去后，Pod 的 CPU 利用率又会降下来，此时对应的 Pod 副本数应该自动减少到一个合理的水平。

CPUUtilizationPercentage 计算过程中使用到的 Pod 的 CPU 使用量通常是 1min 内的平均值，目前通过查询 Heapster 扩展组件来得到这个值，所以需要安装部署 Heapster，这样一来便增加了系统的复杂度和实施 HPA 特性的复杂度，因此，未来的计划是 Kubernetes 自身实现一个基础性能数据采集模块，从而更好地支持 HPA 和其他需要用到基础性能数据的功能模块。此外，我们也看到，如果目标 Pod 没有定义 Pod Request 的值，则无法使用 CPUUtilizationPercentage 来实现 Pod 横向自动扩容的能力。除了使用 CPUUtilizationPercentage，Kubernetes 从 v1.2 版本开始尝试支持应用程序自定义的度量指标，目前仍然为实验特性，不建议在生产环境中使用。

下面是 HPA 定义的一个具体例子：

```
apiVersion: autoscaling/v1
kind: HorizontalPodAutoscaler
metadata:
  name: php-apache
  namespace: default
spec:
  maxReplicas: 10
  minReplicas: 1
  scaleTargetRef:
    kind: Deployment
    name: php-apache
  targetCPUUtilizationPercentage: 90
```

根据上面的定义，我们可以知道这个 HPA 控制的目标对象为一个名叫 php-apache 的 Deployment 里的 Pod 副本，当这些 Pod 副本的 CPUUtilizationPercentage 的值超过 90% 时会触发自动动态扩容行为，扩容或缩容时必须满足的一个约束条件是 Pod 的副本数要介于 1 与 10 之间。

除了可以通过直接定义 yaml 文件并且调用 `kubectl create` 的命令来创建一个 HPA 资源对象的方式，我们还能通过下面的简单命令行直接创建等价的 HPA 对象：

```
# kubectl autoscale deployment php-apache --cpu-percent=90 --min=1 --max=10
```

第 2 章将会给出一个完整的 HPA 例子来说明其用法和功能。

### 1.4.8 StatefulSet

在 Kubernetes 系统中，Pod 的管理对象 RC、Deployment、DaemonSet 和 Job 都是面向无状态的服务。但现实中有很多服务是有状态的，特别是一些复杂的中间件集群，例如 MySQL 集群、MongoDB 集群、Akka 集群、ZooKeeper 集群等，这些应用集群有以下一些共同点。

- ◎ 每个节点都有固定的身份 ID，通过这个 ID，集群中的成员可以相互发现并且通信。
- ◎ 集群的规模是比较固定的，集群规模不能随意变动。

- ◎ 集群里的每个节点都是有状态的，通常会持久化数据到永久存储中。
- ◎ 如果磁盘损坏，则集群里的某个节点无法正常运行，集群功能受损。

如果用 RC/Deployment 控制 Pod 副本数的方式来实现上述有状态的集群，则我们会发现第 1 点是无法满足的，因为 Pod 的名字是随机产生的，Pod 的 IP 地址也是在运行期才确定且可能有变动的，我们事先无法为每个 Pod 确定唯一不变的 ID。另外，为了能够在其他节点上恢复某个失败的节点，这种集群中的 Pod 需要挂接某种共享存储，为了解决这个问题，Kubernetes 从 v1.4 版本开始引入了 PetSet 这个新的资源对象，并且在 v1.5 版本时更名为 StatefulSet，StatefulSet 从本质上来说，可以看作 Deployment/RC 的一个特殊变种，它有如下一些特性。

- ◎ StatefulSet 里的每个 Pod 都有稳定、唯一的网络标识，可以用来发现集群内的其他成员。假设 StatefulSet 的名字叫 kafka，那么第 1 个 Pod 叫 kafka-0，第 2 个叫 kafka-1，以此类推。
- ◎ StatefulSet 控制的 Pod 副本的启停顺序是受控的，操作第  $n$  个 Pod 时，前  $n-1$  个 Pod 已经是运行且准备好的状态。
- ◎ StatefulSet 里的 Pod 采用稳定的持久化存储卷，通过 PV/PVC 来实现，删除 Pod 时默认不会删除与 StatefulSet 相关的存储卷（为了保证数据的安全）。

StatefulSet 除了要与 PV 卷捆绑使用以存储 Pod 的状态数据，还要与 Headless Service 配合使用，即在每个 StatefulSet 的定义中要声明它属于哪个 Headless Service。Headless Service 与普通 Service 的关键区别在于，它没有 Cluster IP，如果解析 Headless Service 的 DNS 域名，则返回的是该 Service 对应的全部 Pod 的 Endpoint 列表。StatefulSet 在 Headless Service 的基础上又为 StatefulSet 控制的每个 Pod 实例创建了一个 DNS 域名，这个域名的格式为：

`$(podname).$(headless service name)`

比如一个 3 节点的 Kafka 的 StatefulSet 集群，对应的 Headless Service 的名字为 kafka，StatefulSet 的名字为 kafka，则 StatefulSet 里面的 3 个 Pod 的 DNS 名称分别为 kafka-0.kafka、kafka-1.kafka、kafka-3.kafka，这些 DNS 名称可以直接在集群的配置文件中固定下来。

## 1.4.9 Service（服务）

### 1. 概述

Service 也是 Kubernetes 里的最核心的资源对象之一，Kubernetes 里的每个 Service 其实就是我们经常提起的微服务架构中的一个“微服务”，之前我们所说的 Pod、RC 等资源对象其实都是为这节所说的“服务”——Kubernetes Service 作“嫁衣”的。图 1.12 显示了 Pod、RC 与 Service 的逻辑关系。



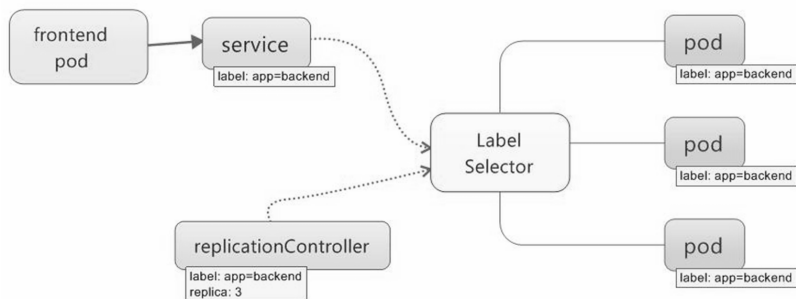


图 1.12 Pod、RC 与 Service 的关系

从图 1.12 中我们看到，Kubernetes 的 Service 定义了一个服务的访问入口地址，前端的应用（Pod）通过这个入口地址访问其背后的一组由 Pod 副本组成的集群实例，Service 与其后端 Pod 副本集群之间则是通过 Label Selector 来实现“无缝对接”的。而 RC 的作用实际上是保证 Service 的服务能力和服务质量始终处于预期的标准。

通过分析、识别并建模系统中的所有服务为微服务——Kubernetes Service，最终我们的系统由多个提供不同业务能力而又彼此独立的微服务单元所组成，服务之间通过 TCP/IP 进行通信，从而形成了我们强大而又灵活的弹性网格，拥有了强大的分布式能力、弹性扩展能力、容错能力，与此同时，我们的程序架构也变得简单和直观许多，如图 1.13 所示。

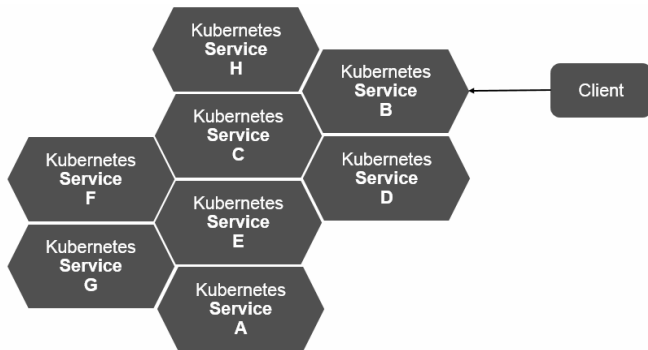


图 1.13 Kubernetes 所提供的微服务网格架构

既然每个 Pod 都会被分配一个单独的 IP 地址，而且每个 Pod 都提供了一个独立的 Endpoint（Pod IP+ContainerPort）以被客户端访问，现在多个 Pod 副本组成了一个集群来提供服务，那么客户端如何来访问它们呢？一般的做法是部署一个负载均衡器（软件或硬件），为这组 Pod 开启一个对外的服务端口如 8000 端口，并且将这些 Pod 的 Endpoint 列表加入 8000 端口的转发列表中，客户端就可以通过负载均衡器的对外 IP 地址+服务端口来访问此服务，而客户端的请求最后会被转发到哪个 Pod，则由负载均衡器的算法所决定。

Kubernetes 也遵循了上述常规做法，运行在每个 Node 上的 kube-proxy 进程其实就是一个智能的软件负载均衡器，它负责把对 Service 的请求转发到后端的某个 Pod 实例上，并在内部实现服务的负载均衡与会话保持机制。但 Kubernetes 发明了一种很巧妙又影响深远的设计：Service 不是共用一个负载均衡器的 IP 地址，而是每个 Service 分配了一个全局唯一的虚拟 IP 地址，这个虚拟 IP 被称为 Cluster IP。这样一来，每个服务就变成了具备唯一 IP 地址的“通信节点”，服务调用就变成了最基础的 TCP 网络通信问题。

我们知道，Pod 的 Endpoint 地址会随着 Pod 的销毁和重新创建而发生改变，因为新 Pod 的 IP 地址与之前旧 Pod 的不同。而 Service 一旦被创建，Kubernetes 就会自动为它分配一个可用的 Cluster IP，而且在 Service 的整个生命周期内，它的 Cluster IP 不会发生改变。于是，服务发现这个棘手的问题在 Kubernetes 的架构里也得以轻松解决：只要用 Service 的 Name 与 Service 的 Cluster IP 地址做一个 DNS 域名映射即可完美解决问题。现在想想，这真是一个很棒的设计。

说了这么久，下面我们动手创建一个 Service，来加深对它的理解。首先我们创建一个名为 tomcat-service.yaml 的定义文件，内容如下：

```
apiVersion: v1
kind: Service
metadata:
  name: tomcat-service
spec:
  ports:
    - port: 8080
  selector:
    tier: frontend
```

上述内容定义了一个名为“tomcat-service”的 Service，它的服务端口为 8080，拥有“tier = frontend”这个 Label 的所有 Pod 实例都属于它，运行下面的命令进行创建：

```
# kubectl create -f tomcat-server.yaml
service "tomcat-service" created
```

注意到我们之前在 tomcat-deployment.yaml 里定义的 Tomcat 的 Pod 刚好拥有这个标签，所以我们刚才创建的 tomcat-service 已经对应到了 Pod 实例，运行下面的命令可以查看 tomcat-service 的 Endpoint 列表，其中 172.17.1.3 是 Pod 的 IP 地址，端口 8080 是 Container 暴露的端口：

```
# kubectl get endpoints
NAME                ENDPOINTS                AGE
kubernetes          192.168.18.131:6443     15d
tomcat-service      172.17.1.3:8080         1m
```

你可能有疑问：“说好的 Service 的 Cluster IP 呢？怎么没有看到？”我们运行下面的命令即可看到 tomcat-service 被分配的 Cluster IP 及更多的信息：

```
# kubectl get svc tomcat-service -o yaml
```

```

apiVersion: v1
kind: Service
metadata:
  creationTimestamp: 2016-07-21T17:05:52Z
  name: tomcat-service
  namespace: default
  resourceVersion: "23964"
  selfLink: /api/v1/namespaces/default/services/tomcat-service
  uid: 61987d3c-4f65-11e6-a9d8-000c29ed42c1
spec:
  clusterIP: 169.169.65.227
  ports:
    - port: 8080
      protocol: TCP
      targetPort: 8080
  selector:
    tier: frontend
  sessionAffinity: None
  type: ClusterIP
status:
  loadBalancer: {}

```

在 `spec.ports` 的定义中, `targetPort` 属性用来确定提供该服务的容器所暴露 (EXPOSE) 的端口号, 即具体业务进程在容器内的 `targetPort` 上提供 TCP/IP 接入; 而 `port` 属性则定义了 Service 的虚端口。前面我们定义 Tomcat 服务时, 没有指定 `targetPort`, 则默认 `targetPort` 与 `port` 相同。

接下来, 我们来看看 Service 的多端口问题。

很多服务都存在多个端口的问题, 通常一个端口提供业务服务, 另外一个端口提供管理服务, 比如 Mycat、Codis 等常见中间件。Kubernetes Service 支持多个 Endpoint, 在存在多个 Endpoint 的情况下, 要求每个 Endpoint 定义一个名字来区分。下面是 Tomcat 多端口的 Service 定义样例:

```

apiVersion: v1
kind: Service
metadata:
  name: tomcat-service
spec:
  ports:
    - port: 8080
      name: service-port
    - port: 8005
      name: shutdown-port
  selector:
    tier: frontend

```

多端口为什么需要给每个端口命名呢? 这就涉及 Kubernetes 的服务发现机制了, 我们接下来进行讲解。

## 2. Kubernetes 的服务发现机制

任何分布式系统都会涉及“服务发现”这个基础问题，大部分分布式系统通过提供特定的 API 接口来实现服务发现的功能，但这样做会导致平台的侵入性比较强，也增加了开发测试的困难。Kubernetes 则采用了直观朴素的思路去解决这个棘手的问题。

首先，每个 Kubernetes 中的 Service 都有一个唯一的 Cluster IP 及唯一的名字，而名字是由开发者自己定义的，部署时也没必要改变，所以完全可以固定在配置中。接下来的问题就是如何通过 Service 的名字找到对应的 Cluster IP？

最早时 Kubernetes 采用了 Linux 环境变量的方式解决这个问题，即每个 Service 生成一些对应的 Linux 环境变量（ENV），并在每个 Pod 的容器在启动时，自动注入这些环境变量，以下是 tomcat-service 产生的环境变量条目：

```
TOMCAT_SERVICE_SERVICE_HOST=169.169.41.218
TOMCAT_SERVICE_SERVICE_PORT_SERVICE_PORT=8080
TOMCAT_SERVICE_SERVICE_PORT_SHUTDOWN_PORT=8005
TOMCAT_SERVICE_SERVICE_PORT=8080
TOMCAT_SERVICE_PORT_8005_TCP_PORT=8005
TOMCAT_SERVICE_PORT=tcp://169.169.41.218:8080
TOMCAT_SERVICE_PORT_8080_TCP_ADDR=169.169.41.218
TOMCAT_SERVICE_PORT_8080_TCP=tcp://169.169.41.218:8080
TOMCAT_SERVICE_PORT_8080_TCP_PROTO=tcp
TOMCAT_SERVICE_PORT_8080_TCP_PORT=8080
TOMCAT_SERVICE_PORT_8005_TCP=tcp://169.169.41.218:8005
TOMCAT_SERVICE_PORT_8005_TCP_ADDR=169.169.41.218
TOMCAT_SERVICE_PORT_8005_TCP_PROTO=tcp
```

上述环境变量中，比较重要的是前 3 条环境变量，我们可以看到，每个 Service 的 IP 地址及端口都是有标准的命名规范的，遵循这个命名规范，就可以通过代码访问系统环境变量的方式得到所需的信息，实现服务调用。

考虑到环境变量的方式获取 Service 的 IP 与端口的方式仍然不太方便，不够直观，后来 Kubernetes 通过 Add-On 增值包的方式引入了 DNS 系统，把服务名作为 DNS 域名，这样一来，程序就可以直接使用服务名来建立通信连接了。目前 Kubernetes 上的大部分应用都已经采用了 DNS 这些新兴的服务发现机制，后面的章节中我们会讲述如何部署这套 DNS 系统。

## 3. 外部系统访问 Service 的问题

为了更加深入地理解和掌握 Kubernetes，我们需要弄明白 Kubernetes 里的“三种 IP”这个关键问题，这三种 IP 分别如下。

- ◎ Node IP: Node 节点的 IP 地址。

- ◎ Pod IP: Pod 的 IP 地址。
- ◎ Cluster IP: Service 的 IP 地址。

首先，Node IP 是 Kubernetes 集群中每个节点的物理网卡的 IP 地址，这是一个真实存在的物理网络，所有属于这个网络的服务器之间都能通过这个网络直接通信，不管它们中是否有部分节点不属于这个 Kubernetes 集群。这也表明了 Kubernetes 集群之外的节点访问 Kubernetes 集群之内的某个节点或者 TCP/IP 服务时，必须要通过 Node IP 进行通信。

其次，Pod IP 是每个 Pod 的 IP 地址，它是 Docker Engine 根据 docker0 网桥的 IP 地址段进行分配的，通常是一个虚拟的二层网络，前面我们说过，Kubernetes 要求位于不同 Node 上的 Pod 能够彼此直接通信，所以 Kubernetes 里一个 Pod 里的容器访问另外一个 Pod 里的容器，就是通过 Pod IP 所在的虚拟二层网络进行通信的，而真实的 TCP/IP 流量则是通过 Node IP 所在的物理网卡流出的。

最后，我们说说 Service 的 Cluster IP，它也是一个虚拟的 IP，但更像是一个“伪造”的 IP 网络，原因有以下几点。

- ◎ Cluster IP 仅仅作用于 Kubernetes Service 这个对象，并由 Kubernetes 管理和分配 IP 地址（来源于 Cluster IP 地址池）。
- ◎ Cluster IP 无法被 Ping，因为没有有一个“实体网络对象”来响应。
- ◎ Cluster IP 只能结合 Service Port 组成一个具体的通信端口，单独的 Cluster IP 不具备 TCP/IP 通信的基础，并且它们属于 Kubernetes 集群这样一个封闭的空间，集群之外的节点如果要访问这个通信端口，则需要做一些额外的工作。
- ◎ 在 Kubernetes 集群之内，Node IP 网、Pod IP 网与 Cluster IP 网之间的通信，采用的是 Kubernetes 自己设计的一种编程方式的特殊的路由规则，与我们所熟知的 IP 路由有很大的不同。

根据上面的分析和总结，我们基本明白了：Service 的 Cluster IP 属于 Kubernetes 集群内部的地址，无法在集群外部直接使用这个地址。那么矛盾来了：实际上我们开发的业务系统中肯定多少有一部分服务是要提供给 Kubernetes 集群外部的应用或者用户来使用的，典型的例子就是 Web 端的服务模块，比如上面的 tomcat-service，那么用户怎么访问它？

采用 NodePort 是解决上述问题的最直接、最有效、最常用的做法。具体做法如下，以 tomcat-service 为例，我们在 Service 的定义里做如下扩展即可（黑体字部分）：

```
apiVersion: v1
kind: Service
metadata:
  name: tomcat-service
```

```
spec:
  type: NodePort
  ports:
    - port: 8080
      nodePort: 31002
  selector:
    tier: frontend
```

其中，nodePort:31002 这个属性表明我们手动指定 tomcat-service 的 NodePort 为 31002，否则 Kubernetes 会自动分配一个可用的端口。接下来，我们在浏览器里访问 http://<nodePort IP>:31002/，就可以看到 Tomcat 的欢迎界面了，如图 1.14 所示。

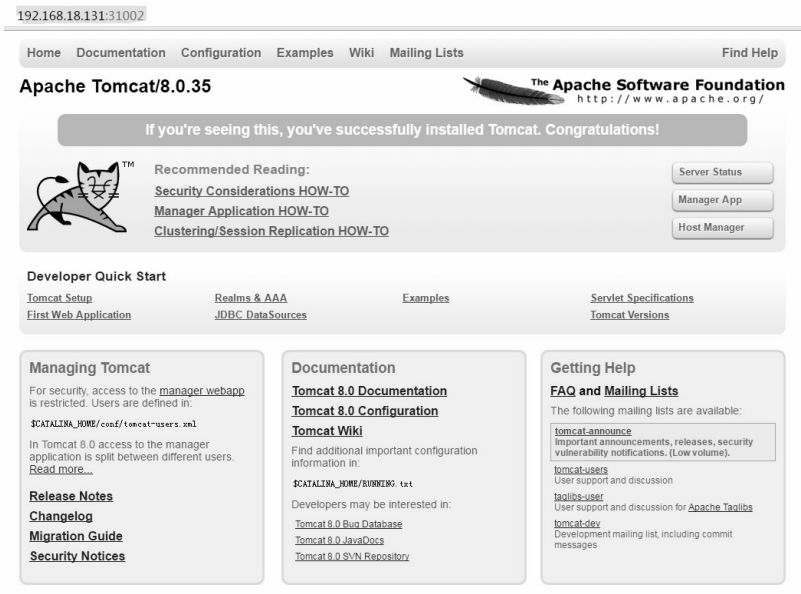


图 1.14 通过 NodePort 访问 Service

NodePort 的实现方式是在 Kubernetes 集群里的每个 Node 上为需要外部访问的 Service 开启一个对应的 TCP 监听端口，外部系统只要用任意一个 Node 的 IP 地址+具体的 NodePort 端口号即可访问此服务，在任意 Node 上运行 netstat 命令，我们就可以看到有 NodePort 端口被监听：

```
# netstat -tlnp | grep 31002
tcp6  0  0  [::]:31002          [::]:*               LISTEN      1125/kube-proxy
```

但 NodePort 还没有完全解决外部访问 Service 的所有问题，比如负载均衡问题，假如我们的集群中有 10 个 Node，则此时最好有一个负载均衡器，外部的请求只需访问此负载均衡器的 IP 地址，由负载均衡器负责转发流量到后面某个 Node 的 NodePort 上。如图 1.15 所示。

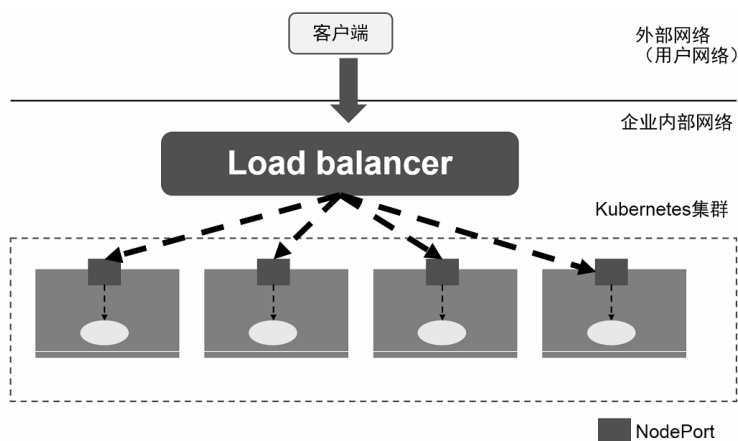


图 1.15 NodePort 与 Load balancer

图 1.15 中的 Load balancer 组件独立于 Kubernetes 集群之外，通常是一个硬件的负载均衡器，或者是以软件方式实现的，例如 HAProxy 或者 Nginx。对于每个 Service，我们通常需要配置一个对应的 Load balancer 实例来转发流量到后端的 Node 上，这的确增加了工作量及出错的概率。于是 Kubernetes 提供了自动化的解决方案，如果我们的集群运行在谷歌的 GCE 公有云上，那么只要我们把 Service 的 `type=NodePort` 改为 `type=LoadBalancer`，此时 Kubernetes 会自动创建一个对应的 Load balancer 实例并返回它的 IP 地址供外部客户端使用。其他公有云提供商只要实现了支持此特性的驱动，则也可以达到上述目的。此外，裸机上的类似机制（Bare Metal Service Load Balancers）也正在被开发。

### 1.4.10 Volume（存储卷）

Volume 是 Pod 中能够被多个容器访问的共享目录。Kubernetes 的 Volume 概念、用途和目的与 Docker 的 Volume 比较类似，但两者不能等价。首先，Kubernetes 中的 Volume 定义在 Pod 上，然后被一个 Pod 里的多个容器挂载到具体的文件目录下；其次，Kubernetes 中的 Volume 与 Pod 的生命周期相同，但与容器的生命周期不相关，当容器终止或者重启时，Volume 中的数据也不会丢失。最后，Kubernetes 支持多种类型的 Volume，例如 GlusterFS、Ceph 等先进的分布式文件系统。

Volume 的使用也比较简单，在大多数情况下，我们先在 Pod 上声明一个 Volume，然后在容器里引用该 Volume 并 Mount 到容器里的某个目录上。举例来说，我们要给之前的 Tomcat Pod 增加一个名字为 `datavol` 的 Volume，并且 Mount 到容器的 `/mydata-data` 目录上，则只要对 Pod 的定义文件做如下修正即可（注意黑体字部分）：

```
template:
  metadata:
    labels:
      app: app-demo
      tier: frontend
  spec:
    volumes:
      - name: datavol
        emptyDir: {}
    containers:
      - name: tomcat-demo
        image: tomcat
        volumeMounts:
          - mountPath: /mydata-data
            name: datavol
        imagePullPolicy: IfNotPresent
```

除了可以让一个 Pod 里的多个容器共享文件、让容器的数据写到宿主机的磁盘上或者写文件到网络存储中，Kubernetes 的 Volume 还扩展出了一种非常有实用价值的功能，即容器配置文件集中化定义与管理，这是通过 ConfigMap 这个新的资源对象来实现的，后面我们会详细说明。

Kubernetes 提供了非常丰富的 Volume 类型，下面逐一进行说明。

## 1. emptyDir

一个 emptyDir Volume 是在 Pod 分配到 Node 时创建的。从它的名称就可以看出，它的初始内容为空，并且无须指定宿主机上对应的目录文件，因为这是 Kubernetes 自动分配的一个目录，当 Pod 从 Node 上移除时，emptyDir 中的数据也会被永久删除。emptyDir 的一些用途如下。

- ◎ 临时空间，例如用于某些应用程序运行时所需的临时目录，且无须永久保留。
- ◎ 长时间任务的中间过程 CheckPoint 的临时保存目录。
- ◎ 一个容器需要从另一个容器中获取数据的目录（多容器共享目录）。

目前，用户无法控制 emptyDir 使用的介质种类。如果 kubelet 的配置是使用硬盘，那么所有 emptyDir 都将创建在该硬盘上。Pod 在将来可以设置 emptyDir 是位于硬盘、固态硬盘上还是基于内存的 tmpfs 上，上面的例子便采用了 emptyDir 类的 Volume。

## 2. hostPath

hostPath 为在 Pod 上挂载宿主主机上的文件或目录，它通常可以用于以下几方面。

- ◎ 容器应用程序生成的日志文件需要永久保存时，可以使用宿主主机的高速文件系统进行存储。



- ◎ 需要访问宿主主机上 Docker 引擎内部数据结构的容器应用时，可以通过定义 `hostPath` 为宿主机 `/var/lib/docker` 目录，使容器内部应用可以直接访问 Docker 的文件系统。

在使用这种类型的 Volume 时，需要注意以下几点。

- ◎ 在不同的 Node 上具有相同配置的 Pod 可能会因为宿主机上的目录和文件不同而导致对 Volume 上目录和文件的访问结果不一致。
- ◎ 如果使用了资源配额管理，则 Kubernetes 无法将 `hostPath` 在宿主机上使用的资源纳入管理。

在下面的例子中使用宿主机的 `/data` 目录定义了一个 `hostPath` 类型的 Volume：

```
volumes:
- name: "persistent-storage"
  hostPath:
    path: "/data"
```

### 3. gcePersistentDisk

使用这种类型的 Volume 表示使用谷歌公有云提供的永久磁盘（Persistent Disk，PD）存放 Volume 的数据，它与 `emptyDir` 不同，PD 上的内容会被永久保存，当 Pod 被删除时，PD 只是被卸载（Unmount），但不会被删除。需要注意的是，你需要先创建一个永久磁盘（PD），才能使用 `gcePersistentDisk`。

使用 `gcePersistentDisk` 有以下一些限制条件。

- ◎ Node（运行 kubelet 的节点）需要是 GCE 虚拟机。
- ◎ 这些虚拟机需要与 PD 存在于相同的 GCE 项目和 Zone 中。

通过 `gcloud` 命令即可创建一个 PD：

```
gcloud compute disks create --size=500GB --zone=us-central1-a my-data-disk
```

定义 `gcePersistentDisk` 类型的 Volume 的示例如下：

```
volumes:
- name: test-volume
  # This GCE PD must already exist.
  gcePersistentDisk:
    pdName: my-data-disk
    fsType: ext4
```

### 4. awsElasticBlockStore

与 GCE 类似，该类型的 Volume 使用亚马逊公有云提供的 EBS Volume 存储数据，需要先创建一个 EBS Volume 才能使用 `awsElasticBlockStore`。

使用 `awsElasticBlockStore` 的一些限制条件如下。

- ◎ Node（运行 kubelet 的节点）需要是 AWS EC2 实例。
- ◎ 这些 AWS EC2 实例需要与 EBS volume 存在于相同的 region 和 availability-zone 中。
- ◎ EBS 只支持单个 EC2 实例 mount 一个 volume。

通过 `aws ec2 create-volume` 命令可以创建一个 EBS volume：

```
aws ec2 create-volume --availability-zone eu-west-1a --size 10 --volume-type gp2
```

定义 `awsElasticBlockStore` 类型的 Volume 的示例如下：

```
volumes:
- name: test-volume
  # This AWS EBS volume must already exist.
  awsElasticBlockStore:
    volumeID: aws://<availability-zone>/<volume-id>
    fsType: ext4
```

## 5. NFS

使用 NFS 网络文件系统提供的共享目录存储数据时，我们需要在系统中部署一个 NFS Server。定义 NFS 类型的 Volume 的示例如下：

```
volumes:
- name: nfs
  nfs:
    # 改为你的 NFS 服务器地址
    server: nfs-server.localhost
    path: "/"
```

## 6. 其他类型的 Volume

- ◎ `iscsi`：使用 iSCSI 存储设备上的目录挂载到 Pod 中。
- ◎ `flocker`：使用 Flocker 来管理存储卷。
- ◎ `glusterfs`：使用开源 GlusterFS 网络文件系统的目录挂载到 Pod 中。
- ◎ `rbd`：使用 Ceph 块设备共享存储（Rados Block Device）挂载到 Pod 中。
- ◎ `gitRepo`：通过挂载一个空目录，并从 GIT 库 clone 一个 git repository 以供 Pod 使用。
- ◎ `secret`：一个 secret volume 用于为 Pod 提供加密的信息，你可以将定义在 Kubernetes 中的 secret 直接挂载为文件让 Pod 访问。secret volume 是通过 `tmfs`（内存文件系统）实现的，所以这种类型的 volume 总是不会持久化的。

### 1.4.11 Persistent Volume

之前我们提到的 Volume 是定义在 Pod 上的，属于“计算资源”的一部分，而实际上，“网络存储”是相对独立于“计算资源”而存在的一种实体资源。比如在使用虚拟机的情况下，我们通常会先定义一个网络存储，然后从中划出一个“网盘”并挂接到虚拟机上。Persistent Volume（简称 PV）和与之相关联的 Persistent Volume Claim（简称 PVC）也起到了类似的作用。

PV 可以理解成 Kubernetes 集群中的某个网络存储中对应的一块存储，它与 Volume 很类似，但有以下区别。

- ◎ PV 只能是网络存储，不属于任何 Node，但可以在每个 Node 上访问。
- ◎ PV 并不是定义在 Pod 上的，而是独立于 Pod 之外定义。
- ◎ PV 目前支持的类型包括：gcePersistentDisk、AWSElasticBlockStore、AzureFile、AzureDisk、FC（Fibre Channel）、Flocker、NFS、iSCSI、RBD（Rados Block Device）、CephFS、Cinder、GlusterFS、VsphereVolume、Quobyte Volumes、VMware Photon、Portworx Volumes、ScaleIO Volumes 和 HostPath（仅供单机测试）。

下面给出了 NFS 类型 PV 的一个 yaml 定义文件，声明了需要 5Gi 的存储空间：

```
apiVersion: v1
kind: PersistentVolume
metadata:
  name: pv0003
spec:
  capacity:
    storage: 5Gi
  accessModes:
    - ReadWriteOnce
  nfs:
    path: /somepath
    server: 172.17.0.2
```

比较重要的是 PV 的 accessModes 属性，目前有以下类型。

- ◎ ReadWriteOnce：读写权限、并且只能被单个 Node 挂载。
- ◎ ReadOnlyMany：只读权限、允许被多个 Node 挂载。
- ◎ ReadWriteMany：读写权限、允许被多个 Node 挂载。

如果某个 Pod 想申请某种类型的 PV，则首先需要定义一个 PersistentVolumeClaim（PVC）对象：

```
kind: PersistentVolumeClaim
apiVersion: v1
```

```
metadata:
  name: myclaim
spec:
  accessModes:
    - ReadWriteOnce
  resources:
    requests:
      storage: 8Gi
```

然后，在 Pod 的 Volume 定义中引用上述 PVC 即可：

```
volumes:
- name: mypd
  persistentVolumeClaim:
    claimName: myclaim
```

最后，我们说说 PV 的状态，PV 是有状态的对象，它有以下几种状态。

- ◎ Available：空闲状态。
- ◎ Bound：已经绑定到某个 PVC 上。
- ◎ Released：对应的 PVC 已经删除，但资源还没有被集群收回。
- ◎ Failed：PV 自动回收失败。

共享存储的原理解析和实践指南详见 3.8 节。

## 1.4.12 Namespace（命名空间）

Namespace（命名空间）是 Kubernetes 系统中的另一个非常重要的概念，Namespace 在很多情况下用于实现多租户的资源隔离。Namespace 通过将集群内部的资源对象“分配”到不同的 Namespace 中，形成逻辑上分组的不同项目、小组或用户组，便于不同的分组在共享使用整个集群的资源的同时还能被分别管理。

Kubernetes 集群在启动后，会创建一个名为“default”的 Namespace，通过 kubectl 可以看到：

```
$ kubectl get namespaces
NAME          LABELS             STATUS
default       <none>             Active
```

接下来，如果不特别指明 Namespace，则用户创建的 Pod、RC、Service 都将被系统创建到这个默认的名为 default 的 Namespace 中。

Namespace 的定义很简单。如下所示的 yaml 定义了名为 development 的 Namespace。

```

apiVersion: v1
kind: Namespace
metadata:
  name: development

```

一旦创建了 `Namespace`，我们在创建资源对象时就可以指定这个资源对象属于哪个 `Namespace`。比如在下面的例子中，我们定义了一个名为 `busybox` 的 `Pod`，放入 `development` 这个 `Namespace` 里：

```

apiVersion: v1
kind: Pod
metadata:
  name: busybox
  namespace: development
spec:
  containers:
  - image: busybox
    command:
    - sleep
    - "3600"
    name: busybox

```

此时，使用 `kubectl get` 命令查看将无法显示：

```

$ kubectl get pods
NAME          READY   STATUS    RESTARTS   AGE

```

这是因为如果不加参数，则 `kubectl get` 命令将仅显示属于“`default`”命名空间的资源对象。

可以在 `kubectl` 命令中加入 `--namespace` 参数来查看某个命名空间中的对象：

```

# kubectl get pods --namespace=development
NAME          READY   STATUS    RESTARTS   AGE
busybox       1/1     Running   0           1m

```

当我们给每个租户创建一个 `Namespace` 来实现多租户的资源隔离时，还能结合 `Kubernetes` 的资源配额管理，限定不同租户能占用的资源，例如 `CPU` 使用量、内存使用量等。关于资源配额管理的问题，在后面的章节中会详细介绍。

### 1.4.13 Annotation（注解）

`Annotation` 与 `Label` 类似，也使用 `key/value` 键值对的形式进行定义。不同的是 `Label` 具有严格的命名规则，它定义的是 `Kubernetes` 对象的元数据（`Metadata`），并且用于 `Label Selector`。而 `Annotation` 则是用户任意定义的“附加”信息，以便于外部工具进行查找，很多时候，`Kubernetes` 的模块自身会通过 `Annotation` 的方式标记资源对象的一些特殊信息。

通常来说，用 Annotation 来记录的信息如下。

- ◎ build 信息、release 信息、Docker 镜像信息等，例如时间戳、release id 号、PR 号、镜像 hash 值、docker registry 地址等。
- ◎ 日志库、监控库、分析库等资源库的地址信息。
- ◎ 程序调试工具信息，例如工具名称、版本号等。
- ◎ 团队的联系信息，例如电话号码、负责人名称、网址等。

#### 1.4.14 小结

上述这些组件是 Kubernetes 系统的核心组件，它们共同构成了 Kubernetes 系统的框架和计算模型。通过对它们进行灵活组合，用户就可以快速、方便地对容器集群进行配置、创建和管理。除了本章所介绍的核心组件，在 Kubernetes 系统中还有许多辅助配置的资源对象，例如 LimitRange、ResourceQuota。另外，一些系统内部使用的对象 Binding、Event 等请参考 Kubernetes 的 API 文档。

在第 2 章中，我们将深入实践并全面掌握 Kubernetes 的各种使用技巧。

# 第 2 章

## Kubernetes 实践指南

---

本章将从 Kubernetes 的系统安装开始，逐步介绍 Kubernetes 的服务相关配置、命令行工具 kubectl 的使用详解，然后通过大量案例实践对 Kubernetes 最核心的容器和微服务架构的概念和用法进行详细说明。

### 2.1 Kubernetes 安装与配置

---

#### 2.1.1 系统要求

---

Kubernetes 系统由一组可执行程序组成，用户可以通过 GitHub 上的 Kubernetes 项目页下载编译好的二进制包，或者下载源代码并编译后进行安装。

安装 Kubernetes 对软件和硬件的系统要求如表 2.1 所示。

表 2.1 安装 Kubernetes 对软件和硬件的系统要求

软 硬 件	最 低 配 置	推 荐 配 置
CPU 和内存	Master: 至少 2core 和 4GB 内存 Node: 至少 4 core 和 16GB 内存	Master: 4 core 和 16GB 内存。 Node: 应根据需要运行的容器数量进行配置
Linux 操作系统	基于 x86_64 架构的各种 Linux 发行版本，包括 Red Hat Linux、CentOS、Fedora、Ubuntu 等，Kernel 版本要求在 3.10 及以上。 也可以在谷歌的 GCE(Google Compute Engine)或者 Amazon 的 AWS (Amazon Web Service) 云平台上进行安装	Red Hat Linux 7 CentOS 7

续表

软 硬 件	最 低 配 置	推 荐 配 置
Docker	1.9 版本及以上 下载和安装说明见 <a href="https://www.docker.com">https://www.docker.com</a>	1.12 版本
etcd	2.0 版本及以上 下载和安装说明见 <a href="https://github.com/coreos/etcd/releases">https://github.com/coreos/etcd/releases</a>	3.0 版本

本章以 CentOS Linux 7 为例，使用 Systemd 系统完成 Kubernetes 服务的配置。其他 Linux 发行版的服务配置请参考相关的系统管理手册。为了便于管理，常见的做法是将 Kubernetes 服务程序配置为 Linux 的系统开机自启动的服务。

需要注意的是，CentOS Linux 7 默认启动了防火墙服务（firewalld），而 Kubernetes 的 Master 与工作 Node 之间会有大量的网络通信，安全的做法是在防火墙上配置各组件需要相互通信的端口号，具体要配置的端口号详见 2.1.6 节中各服务监听的端口号说明。在一个安全的内部网络环境中可以关闭防火墙服务：

```
# systemctl disable firewalld
# systemctl stop firewalld
```

另外，建议在主机上禁用 SELinux，目的是让容器可以读取主机文件系统：

```
# setenforce 0
```

或修改系统文件/etc/sysconfig/selinux，将 SELINUX=enforcing 修改成 SELINUX=disabled，然后重启 Linux。

### 2.1.2 使用 kubeadm 工具快速安装 Kubernetes 集群

最简单的安装方法是使用 yum install kubernetes 命令完成 Kubernetes 集群的安装，但仍需修改各组件的启动参数，才能完成 Kubernetes 集群的配置，整个过程比较复杂，也容易出错，因此从 Kubernetes v1.4 版本开始引入了命令行工具 kubeadm，致力于简化集群的安装和解决 Kubernetes 集群的高可用问题。但直到 v1.6 版本，此工具还不能用于生产环境，但很适合初学者快速安装和学习 Kubernetes。本节我们先学习基于 kubeadm 的安装过程（以 Centos 7 为例），下一节再对以二进制文件手动安装和配置 Kubernetes 集群进行详细说明。

#### 1. 安装 kubeadm 和相关工具

首先配置 yum 源，官方 yum 源地址为 [http://yum.kubernetes.io/repos/kubernetes-el7-x86\\_64](http://yum.kubernetes.io/repos/kubernetes-el7-x86_64)。如果无法访问，则也可以使用国内的一个 yum 源 [https://yumrepo.b0.upaiyun.com/centos/7/x86\\_64](https://yumrepo.b0.upaiyun.com/centos/7/x86_64)，配置文件/etc/yum.repos.d/mritd.repo 的内容如下：



```
[mritd]
name=Mritd Repository
baseurl=https://yumrepo.b0.upaiyun.com/centos/7/x86_64
enabled=1
gpgcheck=1
gpgkey=https://mritd.b0.upaiyun.com/keys/rpm.public.key
```

然后运行 `yum install` 命令安装 `kubeadm` 和相关工具（如果已安装 `Docker`，则取消下面命令中的 `docker`）：

```
# yum install -y docker kubelet kubeadm kubectl kubernetes-cni
```

运行下面的命令，启动 `Docker` 服务与 `kubelet` 服务，并设置为开机自动启动：

```
# systemctl enable docker && systemctl start docker
# systemctl enable kubelet && systemctl start kubelet
```

## 2. 下载 Kubernetes 的相关镜像

由于 `kubeadm` 将自动下载 Kubernetes 的相关镜像，并且默认都是从 `gcr.io` 进行下载的，所以对于无法访问 `gcr.io` 的网络环境，建议修改 `Docker` 的配置文件，增加 `Registry Mirror` 参数，从国内镜像托管站点（例如 `Daocloud`）获取镜像加速服务，将加速码写入配置参数里，即 `OPTIONS='--registry-mirror=http://68e02ab9.m.daocloud.io'`，然后重启 `Docker` 服务。

接下来手动下载 Kubernetes 的相关镜像，下载地址为 `https://hub.docker.com/r/warrior`，下载后将镜像名改为以 `gcr.io/google_container` 开头的名字，以供 `kubeadm` 使用。

```
# docker pull warrior/pause-amd64:3.0
# docker tag warrior/pause-amd64:3.0 gcr.io/google_containers/pause-amd64:3.0

# docker pull warrior/etcd-amd64:3.0.17
# docker tag warrior/etcd-amd64:3.0.17
gcr.io/google_containers/etcd-amd64:3.0.17

# docker pull warrior/kube-apiserver-amd64:v1.6.0
# docker tag warrior/kube-apiserver-amd64:v1.6.0
gcr.io/google_containers/kube-apiserver-amd64:v1.6.0

# docker pull warrior/kube-scheduler-amd64:v1.6.0
# docker tag warrior/kube-scheduler-amd64:v1.6.0
gcr.io/google_containers/kube-scheduler-amd64:v1.6.0

# docker pull warrior/kube-controller-manager-amd64:v1.6.0
# docker tag warrior/kube-controller-manager-amd64:v1.6.0
gcr.io/google_containers/kube-controller-manager-amd64:v1.6.0

# docker pull warrior/kube-proxy-amd64:v1.6.0
```

```
# docker tag warrior/kube-proxy-amd64:v1.6.0
gcr.io/google_containers/kube-proxy-amd64:v1.6.0

# docker pull gysan/dnsmasq-metrics-amd64:1.0
# docker tag gysan/dnsmasq-metrics-amd64:1.0
gcr.io/google_containers/dnsmasq-metrics-amd64:1.0

# docker pull warrior/k8s-dns-kube-dns-amd64:1.14.1
# docker tag warrior/k8s-dns-kube-dns-amd64:1.14.1
gcr.io/google_containers/k8s-dns-kube-dns-amd64:1.14.1

# docker pull warrior/k8s-dns-dnsmasq-nanny-amd64:1.14.1
# docker tag warrior/k8s-dns-dnsmasq-nanny-amd64:1.14.1
gcr.io/google_containers/k8s-dns-dnsmasq-nanny-amd64:1.14.1

# docker pull warrior/k8s-dns-sidecar-amd64:1.14.1
# docker tag warrior/k8s-dns-sidecar-amd64:1.14.1
gcr.io/google_containers/k8s-dns-sidecar-amd64:1.14.1

# docker pull awa305/kube-discovery-amd64:1.0
# docker tag awa305/kube-discovery-amd64:1.0
gcr.io/google_containers/kube-discovery-amd64:1.0

# docker pull gysan/exechealthz-amd64:1.2
# docker tag gysan/exechealthz-amd64:1.2
gcr.io/google_containers/exechealthz-amd64:1.2
```

### 3. 运行 kubeadm init 安装 Master

至此，准备工作已就绪，执行 `kubeadm init` 命令即可一键完成 Kubernetes Master 节点的安装：

```
# kubeadm init --kubernetes-version=1.6.0
```

运行后，控制台将输出如下内容：

```
[kubeadm] WARNING: kubeadm is in beta, please do not use it for production clusters.
[init] Using Kubernetes version: v1.6.4
[init] Using Authorization mode: RBAC
[preflight] Running pre-flight checks
[preflight] Starting the kubelet service
[certificates] Generated CA certificate and key.
[certificates] Generated API server certificate and key.
[certificates] API Server serving cert is signed for DNS names .....
[certificates] Generated API server kubelet client certificate and key.
[certificates] Generated service account token signing key and public key.
[certificates] Generated front-proxy CA certificate and key.
```

```
[certificates] Generated front-proxy client certificate and key.
[certificates] Valid certificates and keys now exist in "/etc/kubernetes/pki"
[kubeconfig] Wrote KubeConfig file to disk: "/etc/kubernetes/scheduler.conf"
[kubeconfig] Wrote KubeConfig file to disk: "/etc/kubernetes/admin.conf"
[kubeconfig] Wrote KubeConfig file to disk: "/etc/kubernetes/kubelet.conf"
[kubeconfig] Wrote KubeConfig file to disk:
"/etc/kubernetes/controller-manager.conf"
[apiclient] Created API client, waiting for the control plane to become ready
```

等待一段时间后，Kubernetes Master 节点安装成功，显示如下信息：

```
[apiclient] All control plane components are healthy after 21.976107 seconds
[apiclient] Waiting for at least one node to register
[apiclient] First node has registered after 7.226536 seconds
[token] Using token: c19078.ba4b5ac7c40cb58e
[apiconfig] Created RBAC rules
[addons] Created essential addon: kube-proxy
[addons] Created essential addon: kube-dns
Your Kubernetes master has initialized successfully!
To start using your cluster, you need to run (as a regular user):
  sudo cp /etc/kubernetes/admin.conf $HOME/
  sudo chown $(id -u):$(id -g) $HOME/admin.conf
  export KUBECONFIG=$HOME/admin.conf
You should now deploy a pod network to the cluster.
Run "kubectl apply -f [podnetwork].yaml" with one of the options listed at:
  http://kubernetes.io/docs/admin/addons/
You can now join any number of machines by running the following on each node
as root:
kubeadm join --token c19078.ba4b5ac7c40cb58e 192.168.18.4:6443
```

按照提示，执行下面的命令复制配置文件到普通用户的 home 目录下：

```
# sudo cp /etc/kubernetes/admin.conf $HOME/
# sudo chown $(id -u):$(id -g) $HOME/admin.conf
# export KUBECONFIG=$HOME/admin.conf
```

至此完成了 Master 节点上 Kubernetes 软件的安装，但集群内还没有可用的工作 Node，并缺乏容器网络的配置。

## 4. 安装 Node，加入集群

对于新节点的添加，在 Node 主机上执行下面的安装过程。

(1) 安装 kubeadm 和相关工具（如果已安装 Docker，则取消下面命令中的 docker）：

```
# yum install -y docker kubelet kubeadm kubectl kubernetes-cni
```

运行下面的命令启动 Docker 服务与 kubelet 服务，并设置为开机自动启动：

```
# systemctl enable docker && systemctl start docker
```

```
# systemctl enable kubelet && systemctl start kubelet
```

(2) 执行 `kubeadm join` 命令，加入集群：

```
# kubeadm join --token c19078.ba4b5ac7c40cb58e 192.168.18.4:6443
```

其中 `token` 值来源于使用 `kubeadm` 安装 Master 过程中提示的最后一行字：

```
You can now join any number of machines by running the following on each node
as root:
```

```
kubeadm join --token c19078.ba4b5ac7c40cb58e 192.168.18.4:6443
```

192.168.18.4:6443 是 Master 的 URL 地址。

`kubeadm` 在 Master 节点也安装了 `kubelet`，在默认情况下并不参与工作负载。如果希望安装一个单机 All-In-One 的 Kubernetes 环境，则可以执行下面的命令（删除 Node 的 Label “node-role.kubernetes.io/master”），让 Master 节点成为一个 Node 节点：

```
# kubectl taint nodes --all node-role.kubernetes.io/master-
```

## 5. 安装网络插件

通过 `kubectl get nodes` 命令，也会发现 Kubernetes 提示 Master 节点为 `NotReady` 状态，这是因为还没有安装 CNI 网络插件：

```
root@kub-master ~]# kubectl get nodes
NAME          STATUS    AGE      VERSION
kub-master    NotReady  31m      v1.6.4
```

根据 `kubeadm` 的提示进行 CNI 网络插件的安装。网络插件有许多选择，可以参考 <http://kubernetes.io/docs/admin/addons/> 的说明。

```
You should now deploy a pod network to the cluster.
Run "kubectl apply -f [podnetwork].yaml" with one of the options listed at:
http://kubernetes.io/docs/admin/addons/
```

选择 `weave` 插件，执行下面的命令即可一键完成安装：

```
# kubectl apply -f https://git.io/weave-kube-1.6
clusterrole "weave-net" created
serviceaccount "weave-net" created
clusterrolebinding "weave-net" created
daemonset "weave-net" created
```

## 6. 验证 Kubernetes 集群安装完成

执行下面的命令，验证 Kubernetes 集群的相关 Pod 是否都正常创建并运行：

```
# kubectl get pods --all-namespaces
NAMESPACE      NAME                                     READY    STATUS    RESTARTS   AGE
```

```

kube-system    etcd-kub-master          1/1      Running    0          43m
kube-system    kube-apiserver-kub-master 1/1      Running    0          43m
kube-system    kube-controller-manager-kub-master 1/1      Running    0          43m
kube-system    kube-dns-3913472980-2fxgp 3/3      Running    0          48m
kube-system    kube-proxy-7x9j2         1/1      Running    0          48m
kube-system    kube-scheduler-kub-master 1/1      Running    0          43m
kube-system    weave-net-4j0qc          2/2      Running    0          13m

```

如果发现有状态错误的 Pod，则可以执行 `kubecttl --namespace=kube-system describe pod <pod_name>` 来查看错误原因，一个常见的原因是 Image 镜像没有下载下来。

至此，通过 `kubeadm` 工具就实现了 Kubernetes 集群的快速搭建。如果安装失败，则可以执行 `kubeadm reset` 命令将主机恢复原状，重新执行 `kubeadm init` 命令再次进行安装。

### 2.1.3 以二进制文件方式安装 Kubernetes 集群

本节以二进制文件方式安装 Kubernetes 集群，并对每个组件的配置进行详细说明。

从 Kubernetes 发布官网 <https://github.com/kubernetes/kubernetes/releases> 找到对应的版本号，单击 CHANGELOG，找到已编译好的二进制文件的下载页面，如图 2.1 所示。本书基于 Kubernetes v1.6 版本进行说明。



v1.6.0

enisoc released this on 29 Mar · 116 commits to release-1.6 since this release

See `kubernetes-announce@` and CHANGELOG for details.

SHA256 for `kubernet.es.tar.gz`: `e89318b88ea340e68c427d0aad701e544ce2291195dc1d5901222e7bae48f03b`

Additional binary downloads are linked in the CHANGELOG.

#### Downloads

- `kubernet.es.tar.gz` 5.8 MB
- Source code (zip)
- Source code (tar.gz)

#### Downloads for v1.6.0

filename	sha256 hash
kubernet.es.tar.gz	e89318b88ea340e68c427d0aad701e544ce2291195dc1d5901222e7bae48f03b
kubernet.es-src.tar.gz	0b03d27e1c7af3be5d94ecd5f679e5f55588d801376cf1ae170d9ec0a229b1e2

图 2.1 GitHub 上 Kubernetes 的下载页面

Server Binaries

filename	sha256 hash
kubernetes-server-linux-amd64.tar.gz	3625b63d573aa4d28eaa30b291017f775f2ddc0523f40d25023ad1da9c30390e
kubernetes-server-linux-arm64.tar.gz	906496c985d4d836466b73e1c9e618ea8ce07f396aba3a96edcdc6045e0ab4e3
kubernetes-server-linux-arm.tar.gz	3b63f481156f57729bc8100566d8b3d7856791e5b36bb70042e17d21f11f8d5d
kubernetes-server-linux-ppc64le.tar.gz	382666b3892fd4d89be5e4bb052dd0ef0d1c1d213c1ae7a92435083a105064fd
kubernetes-server-linux-s390x.tar.gz	e15de8896bd84a9b74756adc1a2e20c6912a65f6ff0a3f3dfabc8b463e31d19b

图 2.1 GitHub 上 Kubernetes 的下载页面（附）

压缩包 `kubernetes.tar.gz` 内包含了 Kubernetes 的服务程序文件、文档和示例；压缩包 `kubernetes-src.tar.gz` 内则包含了全部源代码。也可以直接下载 `Server Binaries` 中的 `kubernetes-server-linux-amd64.tar.gz` 文件,其中包含了 Kubernetes 需要运行的全部服务程序文件。主要的服务程序文件列表如表 2.2 所示。

表 2.2 主要的服务程序文件列表

文 件 名	说 明
kube-apiserver	apiserver 主程序
kube-apiserver.docker_tag	apiserver docker 镜像的 tag
kube-apiserver.tar	apiserver docker 镜像文件
kube-controller-manager	controller-manager 主程序
kube-controller-manager.docker_tag	controller-manager docker 镜像的 tag
kube-controller-manager.tar	controller-manager docker 镜像文件
kube-scheduler	scheduler 主程序
kube-scheduler.docker_tag	scheduler docker 镜像的 tag
kube-scheduler.tar	scheduler docker 镜像文件
kubelet	kubelet 主程序
kube-proxy	proxy 主程序
kubectrl	客户端命令行工具
kubeadm	Kubernetes 集群安装的命令行工具
hyperkube	包含了所有服务的程序，可以启动任一服务
cloud-controller-manager	Alpha 版特性，v1.6 版引入，提供与云提供商服务对接的各 Controller
kube-aggregator	Alpha 版特性，v1.6 版引入，为集群提供应用定制性能指标，用于 HPA

Kubernetes 的服务都可以通过直接运行二进制文件加上启动参数完成。在 Kubernetes Master 节点上需要部署 `etcd`、`kube-apiserver`、`kube-controller-manager`、`kube-scheduler` 服务进程，在工作 Node 上需要部署 `docker`、`kubelet` 和 `kube-proxy` 服务进程。

将 Kubernetes 的二进制可执行文件复制到 `/usr/bin`（如果复制到其他目录，则相应地将

systemd 服务文件中的文件路径修改正确即可), 就完成了软件的安装。若要使 Kubernetes 正常工作, 则主要的工作是对各个服务进行详细配置。

下面主要介绍最重要的服务启动参数, 每个服务的启动参数还有很多, 详见 2.1.6 节的完整说明。读者可以尝试启用和修改参数的值, 以观察服务运行的不同效果。

## 1. Master 上的 etcd、kube-apiserver、kube-controller-manager、kube-scheduler 服务

### 1) etcd 服务

etcd 服务作为 Kubernetes 集群的主数据库, 在安装 Kubernetes 各服务之前需要首先安装和启动。

从 GitHub 官网(<https://github.com/coreos/etcd/releases>)下载 etcd 二进制文件, 将 etcd 和 etcdctl 文件复制到/usr/bin 目录。

设置 systemd 服务文件/usr/lib/systemd/system/etcd.service:

```
[Unit]
Description=Etcd Server
After=network.target

[Service]
Type=simple
WorkingDirectory=/var/lib/etcd/
EnvironmentFile=-/etc/etcd/etcd.conf
ExecStart=/usr/bin/etcd

[Install]
WantedBy=multi-user.target
```

其中 WorkingDirectory (/var/lib/etcd/) 表示 etcd 数据保存的目录, 需要在启动 etcd 服务之前进行创建。

配置文件/etc/etcd/etcd.conf 通常不需要特别的参数设置 (详细的参数配置内容参见官方文档), etcd 默认将监听在 http://127.0.0.1:2379 地址供客户端连接。

配置完成后, 通过 systemctl start 命令启动 etcd 服务。同时, 使用 systemctl enable 命令将服务加入开机启动列表中:

```
# systemctl daemon-reload
# systemctl enable etcd.service
# systemctl start etcd.service
```

通过执行 etcdctl cluster-health, 可以验证 etcd 是否正确启动:

```
# etcdctl endpoint health
```

```
127.0.0.1:2379 is healthy: successfully committed proposal: took = 1.33112ms
```

## 2) kube-apiserver 服务

将 kube-apiserver、kube-controller-manager 和 kube-scheduler 文件复制到/usr/bin 目录。

编辑 systemd 服务文件/usr/lib/systemd/system/kube-apiserver.service，内容如下：

```
[Unit]
Description=Kubernetes API Server
Documentation=https://github.com/GoogleCloudPlatform/kubernetes
After=etcd.service
Wants=etcd.service

[Service]
EnvironmentFile=/etc/kubernetes/apiserver
ExecStart=/usr/bin/kube-apiserver $KUBE_API_ARGS
Restart=on-failure
Type=notify
LimitNOFILE=65536

[Install]
WantedBy=multi-user.target
```

配置文件/etc/kubernetes/apiserver 的内容包括了 kube-apiserver 的全部启动参数，主要的配置参数在变量 KUBE\_API\_ARGS 中指定。

```
# cat /etc/kubernetes/apiserver
KUBE_API_ARGS="--storage-backend=etcd3 --etcd-servers=http://127.0.0.1:2379
--insecure-bind-address=0.0.0.0 --insecure-port=8080
--service-cluster-ip-range=169.169.0.0/16 --service-node-port-range=1-65535
--admission_control=NamespaceLifecycle,LimitRanger,ServiceAccount,DefaultStorage
Class,ResourceQuota --logtostderr=false --log-dir=/var/log/kubernetes --v=2"
```

对启动参数的说明如下。

- ◎ --etcd-servers: 指定 etcd 服务的 URL。
- ◎ --storage-backend: 指定 etcd 的版本，从 Kubernetes v1.6 开始，默认为 etcd3。注意，在 Kubernetes v1.6 之前的版本中没有这个参数，kube-apiserver 默认使用 etcd2，对于正在运行的 v1.5 或旧版本的 Kubernetes 集群，etcd 提供了数据升级方案，详见 etcd 文档（[https://coreos.com/etcd/docs/latest/upgrades/upgrade\\_3\\_0.html](https://coreos.com/etcd/docs/latest/upgrades/upgrade_3_0.html)）。
- ◎ --insecure-bind-address: apiserver 绑定主机的非安全 IP 地址，设置 0.0.0.0 表示绑定所有 IP 地址。
- ◎ --insecure-port: apiserver 绑定主机的非安全端口号，默认为 8080。
- ◎ --service-cluster-ip-range: Kubernetes 集群中 Service 的虚拟 IP 地址段范围，以 CIDR 格



式表示，例如 169.169.0.0/16，该 IP 范围不能与物理机的真实 IP 段有重合。

- ◎ **--service-node-port-range**: Kubernetes 集群中 Service 可映射的物理机端口号范围，默认为 30000~32767。
- ◎ **--admission\_control**: Kubernetes 集群的准入控制设置，各控制模块以插件的形式依次生效。
- ◎ **--logtostderr**: 设置为 false 表示将日志写入文件，不写入 stderr。
- ◎ **--log-dir**: 日志目录。
- ◎ **--v**: 日志级别。

### 3) kube-controller-manager 服务

kube-controller-manager 服务依赖于 kube-apiserver 服务：

```
# cat /usr/lib/systemd/system/kube-controller-manager.service
[Unit]
Description=Kubernetes Controller Manager
Documentation=https://github.com/GoogleCloudPlatform/kubernetes
After=kube-apiserver.service
Requires=kube-apiserver.service

[Service]
EnvironmentFile=/etc/kubernetes/controller-manager
ExecStart=/usr/bin/kube-controller-manager $KUBE_CONTROLLER_MANAGER_ARGS
Restart=on-failure
LimitNOFILE=65536

[Install]
WantedBy=multi-user.target
```

配置文件/etc/kubernetes/controller-manager 的内容包含了 kube-controller-manager 的全部启动参数，主要的配置参数在变量 KUBE\_CONTROLLER\_MANAGER\_ARGS 中指定：

```
# cat /etc/kubernetes/controller-manager
KUBE_CONTROLLER_MANAGER_ARGS="--master=http://192.168.18.3:8080
--logtostderr=false --log-dir=/var/log/kubernetes --v=2"
```

对启动参数的说明如下。

- ◎ **--master**: 指定 apiserver 的 URL 地址。
- ◎ **--logtostderr**: 设置为 false 表示将日志写入文件，不写入 stderr。
- ◎ **--log-dir**: 日志目录。
- ◎ **--v**: 日志级别。

#### 4) kube-scheduler 服务

kube-scheduler 服务也依赖于 kube-apiserver 服务：

```
# cat /usr/lib/systemd/system/kube-controller-manager.service
[Unit]
Description=Kubernetes Controller Manager
Documentation=https://github.com/GoogleCloudPlatform/kubernetes
After=kube-apiserver.service
Requires=kube-apiserver.service

[Service]
EnvironmentFile=/etc/kubernetes/scheduler
ExecStart=/usr/bin/kube-scheduler $KUBE_SCHEDULER_ARGS
Restart=on-failure
LimitNOFILE=65536

[Install]
WantedBy=multi-user.target
```

配置文件/etc/kubernetes/scheduler 的内容包括了 kube-scheduler 的全部启动参数，主要的配置参数在变量 KUBE\_SCHEDULER\_ARGS 中指定：

```
# cat /etc/kubernetes/scheduler

KUBE_SCHEDULER_ARGS="--master=http://192.168.18.3:8080 --logtostderr=false
--log-dir=/var/log/kubernetes --v=2"
```

对启动参数的说明如下。

- ◎ --master: 指定 apiserver 的 URL 地址。
- ◎ --logtostderr: 设置为 false 表示将日志写入文件，不写入 stderr。
- ◎ --log-dir: 日志目录。
- ◎ --v: 日志级别。

配置完成后，执行 systemctl start 命令按顺序启动这 3 个服务，同时，使用 systemctl enable 命令将服务加入开机启动列表中：

```
# systemctl daemon-reload
# systemctl enable kube-apiserver.service
# systemctl start kube-apiserver.service
# systemctl enable kube-controller-manager
# systemctl start kube-controller-manager
# systemctl enable kube-scheduler
# systemctl start kube-scheduler
```

通过 systemctl status <service\_name> 来验证服务的启动状态，“running”表示启动成功。

至此，Master 上所需的服务就全部启动完成了。

## 2. Node 上的 kubelet、kube-proxy 服务

在工作 Node 节点上需要预先安装好 Docker Daemon 并且正常启动。Docker 的安装和启动详见 Docker 官网 <http://www.docker.com> 的说明文档。

### 1) kubelet 服务

kubelet 服务依赖于 Docker 服务。

```
# cat /usr/lib/systemd/system/kubelet.service
[Unit]
Description=Kubernetes Kubelet Server
Documentation=https://github.com/GoogleCloudPlatform/kubernetes
After=docker.service
Requires=docker.service

[Service]
WorkingDirectory=/var/lib/kubelet
EnvironmentFile=/etc/kubernetes/kubelet
ExecStart=/usr/bin/kubelet $KUBELET_ARGS
Restart=on-failure

[Install]
WantedBy=multi-user.target
```

其中 WorkingDirectory 表示 kubelet 保存数据的目录，需要在启动 kubelet 服务之前进行创建。

配置文件/etc/kubernetes/kubelet 的内容包括了 kubelet 的全部启动参数，主要的配置参数在变量 KUBELET\_ARGS 中指定。

```
# cat /etc/kubernetes/kubelet
KUBELET_ARGS="--api-servers=http://192.168.18.3:8080
--hostname-override=192.168.18.3 --logtostderr=false
--log-dir=/var/log/kubernetes --v=2"
```

对启动参数的说明如下。

- ◎ --api-servers: 指定 apiserver 的 URL 地址，可以指定多个。
- ◎ --hostname-override: 设置本 Node 的名称。
- ◎ --logtostderr: 设置为 false 表示将日志写入文件，不写入 stderr。
- ◎ --log-dir: 日志目录。
- ◎ --v: 日志级别。

## 2) kube-proxy 服务

kube-proxy 服务依赖于 network 服务：

```
[Unit]
Description=Kubernetes Kube-Proxy Server
Documentation=https://github.com/GoogleCloudPlatform/kubernetes
After=network.target
Requires=network.service

[Service]
EnvironmentFile=/etc/kubernetes/proxy
ExecStart=/usr/bin/kube-proxy $KUBE_PROXY_ARGS
Restart=on-failure
LimitNOFILE=65536

[Install]
WantedBy=multi-user.target
```

配置文件/etc/kubernetes/proxy 的内容包括了 kube-proxy 的全部启动参数，主要的配置参数在变量 KUBE\_PROXY\_ARGS 中指定：

```
# cat /etc/kubernetes/proxy
KUBE_PROXY_ARGS="--master=http://192.168.18.3:8080 --logtostderr=false
--log-dir=/var/log/kubernetes --v=2"
```

对启动参数的说明如下。

- ◎ --master: 指定 apiserver 的 URL 地址。
- ◎ --logtostderr: 设置为 false 表示将日志写入文件，不写入 stderr。
- ◎ --log-dir: 日志目录。
- ◎ --v: 日志级别。

配置完成后，通过 systemctl 启动 kubelet 和 kube-proxy 服务：

```
# systemctl daemon-reload
# systemctl enable kubelet.service
# systemctl start kubelet.service
# systemctl enable kube-proxy
# systemctl start kube-proxy
```

kubelet 默认采用向 Master 自动注册本 Node 的机制，在 Master 上查看各 Node 的状态，状态为 Ready 表示 Node 已经成功注册并且状态为可用：

```
# kubectl get nodes
NAME                STATUS    AGE
192.168.18.3        Ready    1m
```

等所有 Node 的状态都为 Ready 之后，一个 Kubernetes 集群就启动完成了。接下来就可以创建 Pod、RC、Service 等资源对象来部署 Docker 容器应用了。

## 2.1.4 Kubernetes 集群的安全设置

### 1. 基于 CA 签名的双向数字证书认证方式

在一个安全的内网环境中，Kubernetes 的各个组件与 Master 之间可以通过 apiserver 的非安全端口 `http://apiserver:8080` 进行访问。但如果 apiserver 需要对外提供服务，或者集群中的某些容器也需要访问 apiserver 以获取集群中的某些信息，则更安全的做法是启用 HTTPS 安全机制。Kubernetes 提供了基于 CA 签名的双向数字证书认证方式和简单的基于 HTTP BASE 或 TOKEN 的认证方式，其中 CA 证书方式的安全性最高。本节先介绍以 CA 证书的方式配置 Kubernetes 集群，要求 Master 上的 kube-apiserver、kube-controller-manager、kube-scheduler 进程及各 Node 上的 kubelet、kube-proxy 进程进行 CA 签名双向数字证书安全设置。

基于 CA 签名的双向数字证书的生成过程如下。

(1) 为 kube-apiserver 生成一个数字证书，并用 CA 证书进行签名。

(2) 为 kube-apiserver 进程配置证书相关的启动参数，包括 CA 证书（用于验证客户端证书的签名真伪）、自己的经过 CA 签名后的证书及私钥。

(3) 为每个访问 Kubernetes API Server 的客户端（如 kube-controller-manager、kube-scheduler、kubelet、kube-proxy 及调用 API Server 的客户端程序 kubectl 等）进程生成自己的数字证书，也都用 CA 证书进行签名，在相关程序的启动参数里增加 CA 证书、自己的证书等相关参数。

#### 1) 设置 kube-apiserver 的 CA 证书相关的文件和启动参数

使用 OpenSSL 工具在 Master 服务器上创建 CA 证书和私钥相关的文件：

```
# openssl genrsa -out ca.key 2048
# openssl req -x509 -new -nodes -key ca.key -subj "/CN=k8s-master" -days 5000
-out ca.crt
# openssl genrsa -out server.key 2048
```

注意：生成 ca.crt 时，-subj 参数中“/CN”的值为 Master 主机名。

准备 master\_ssl.cnf 文件，该文件用于 x509 v3 版本的证书。在该文件中主要需要设置 Master 服务器的 hostname（k8s-master）、IP 地址（192.168.18.3），以及 Kubernetes Master Service 的虚拟服务名称（kubernetes.default 等）和该虚拟服务的 ClusterIP 地址（169.169.0.1）。

master\_ssl.cnf 文件的示例如下：

```
[req]
```

```
req_extensions = v3_req
distinguished_name = req_distinguished_name
[req_distinguished_name]
[ v3_req ]
basicConstraints = CA:FALSE
keyUsage = nonRepudiation, digitalSignature, keyEncipherment
subjectAltName = @alt_names
[alt_names]
DNS.1 = kubernetes
DNS.2 = kubernetes.default
DNS.3 = kubernetes.default.svc
DNS.4 = kubernetes.default.svc.cluster.local
DNS.5 = k8s-master
IP.1 = 169.169.0.1
IP.2 = 192.168.18.3
```

基于 master\_ssl.cnf 创建 server.csr 和 server.crt 文件。在生成 server.csr 时，-subj 参数中“/CN”的值需为 Master 的主机名：

```
# openssl req -new -key server.key -subj "/CN=k8s-master" -config master_ssl.cnf
-out server.csr
# openssl x509 -req -in server.csr -CA ca.crt -CAkey ca.key -CAcreateserial -days
5000 -extensions v3_req -extfile master_ssl.cnf -out server.crt
```

全部执行完后会生成 6 个文件：ca.crt、ca.key、ca.srl、server.crt、server.csr、server.key。

将这些文件复制到一个目录中（例如/var/run/kubernetes/），然后设置 kube-apiserver 的三个启动参数 “--client-ca-file” “--tls-cert-file” 和 “--tls-private-key-file”，分别代表 CA 根证书文件、服务端证书文件和服务端私钥文件：

```
--client-ca-file=/var/run/kubernetes/ca.crt
--tls-private-key-file=/var/run/kubernetes/server.key
--tls-cert-file=/var/run/kubernetes/server.crt
```

同时，可以关掉非安全端口 8080，设置安全端口为 6443（默认值）：

```
--insecure-port=0
--secure-port=6443
```

最后重启 kube-apiserver 服务。

## 2) 设置 kube-controller-manager 的客户端证书、私钥和启动参数

```
$ openssl genrsa -out cs_client.key 2048
$ openssl req -new -key cs_client.key -subj "/CN=k8s-master" -out cs_client.csr
$ openssl x509 -req -in cs_client.csr -CA ca.crt -CAkey ca.key -CAcreateserial
-out cs_client.crt -days 5000
```

其中，在生成 cs\_client.crt 时，-CA 参数和-CAkey 参数使用的是 apiserver 的 ca.crt 和 ca.key 文件。然后将这些文件复制到一个目录中（例如/var/run/kubernetes/）。

接下来创建/etc/kubernetes/kubeconfig 文件(kube-controller-manager 与 kube-scheduler 共用), 配置客户端证书等相关参数, 内容如下:

```
apiVersion: v1
kind: Config
users:
- name: controllermanager
  user:
    client-certificate: /var/run/kubernetes/cs_client.crt
    client-key: /var/run/kubernetes/cs_client.key
clusters:
- name: local
  cluster:
    certificate-authority: /var/run/kubernetes/ca.crt
contexts:
- context:
    cluster: local
    user: controllermanager
  name: my-context
current-context: my-context
```

然后, 设置 kube-controller-manager 服务的启动参数, 注意, --master 的地址为 HTTPS 安全服务地址, 不使用非安全地址 http://192.168.18.3:8080:

```
--master=https://192.168.18.3:6443
--service-account-key-file=/var/run/kubernetes/server.key
--root-ca-file=/var/run/kubernetes/ca.crt
--kubeconfig=/etc/kubernetes/kubeconfig
```

重启 kube-controller-manager 服务。

### 3) 设置 kube-scheduler 启动参数

kube-scheduler 复用上一步 kube-controller-manager 创建的客户端证书, 配置启动参数:

```
--master=https://192.168.18.3:6443
--kubeconfig=/etc/kubernetes/kubeconfig
```

重启 kube-scheduler 服务。

### 4) 设置每台 Node 上 kubelet 的客户端证书、私钥和启动参数

首先复制 kube-apiserver 的 ca.crt 和 ca.key 文件到 Node 上, 在生成 kubelet\_client.crt 时-CA 参数和-CAkey 参数使用的是 apiserver 的 ca.crt 和 ca.key 文件;在生成 kubelet\_client.csr 时将-subj 参数中的 “/CN” 设置为本 Node 的 IP 地址:

```
$ openssl genrsa -out kubelet_client.key 2048
$ openssl req -new -key kubelet_client.key -subj "/CN=192.168.18.4" -out
kubelet_client.csr
$ openssl x509 -req -in kubelet_client.csr -CA ca.crt -CAkey ca.key
```

```
-CAcreateserial -out kubelet_client.crt -days 5000
```

将这些文件复制到一个目录中（例如/var/run/kubernetes/）。

接下来创建/etc/kubernetes/kubeconfig 文件（kubelet 和 kube-proxy 进程共用），配置客户端证书等相关参数，内容如下：

```
apiVersion: v1
kind: Config
users:
- name: kubelet
  user:
    client-certificate: /etc/kubernetes/ssl_keys/kubelet_client.crt
    client-key: /etc/kubernetes/ssl_keys/kubelet_client.key
clusters:
- name: local
  cluster:
    certificate-authority: /etc/kubernetes/ssl_keys/ca.crt
contexts:
- context:
    cluster: local
    user: kubelet
  name: my-context
current-context: my-context
```

然后，设置 kubelet 服务的启动参数：

```
--api-servers=https://192.168.18.3:6443
--kubeconfig=/etc/kubernetes/kubeconfig
```

最后重启 kubelet 服务。

## 5) 设置 kube-proxy 的启动参数

kube-proxy 复用上一步 kubelet 创建的客户端证书，配置启动参数：

```
--master=https://192.168.18.3:6443
--kubeconfig=/etc/kubernetes/kubeconfig
```

重启 kube-proxy 服务。

至此，一个基于 CA 的双向数字证书认证的 Kubernetes 集群环境就搭建完成了。

## 6) 设置 kubectl 客户端使用安全方式访问 apiserver

在使用 kubectl 对 Kubernetes 集群进行操作时，默认使用非安全端口 8080 对 apiserver 进行访问，也可以设置为安全访问 apiserver 的模式，需要设置 3 个证书相关的参数“--certificate-authority”“--client-certificate”和“--client-key”，分别表示用于 CA 授权的证书、客户端证书和客户端密钥。



- ◎ `--certificate-authority`: 使用为 kube-apiserver 生成的 `ca.crt` 文件。
- ◎ `--client-certificate`: 使用为 kube-controller-manager 生成的 `cs_client.crt` 文件。
- ◎ `--client-key`: 使用为 kube-controller-manager 生成的 `cs_client.key` 文件。

同时, 指定 apiserver 的 URL 地址为 HTTPS 安全地址 (例如 `https://k8s-master:443`), 最后输入需要执行的子命令, 即可对 apiserver 进行安全访问了:

```
# kubectl --server=https://192.168.18.3:6443
--certificate-authority=/etc/kubernetes/ssl_keys/ca.crt
--client-certificate=/etc/kubernetes/ssl_keys/cs_client.crt
--client-key=/etc/kubernetes/ssl_keys/cs_client.key get nodes
```

NAME	STATUS	AGE
k8s-node-1	Ready	1h

## 2. 基于 HTTP BASE 或 TOKEN 的简单认证方式

除了基于 CA 的双向数字证书认证方式, Kubernetes 也提供了基于 HTTP BASE 或 TOKEN 的简单认证方式。各组件与 apiserver 之间的通信方式仍然采用 HTTPS, 但不使用 CA 数字证书。

采用基于 HTTP BASE 或 TOKEN 的简单认证方式时, API Server 对外暴露 HTTPS 端口, 客户端提供用户名、密码或 Token 来完成认证过程。需要说明的是, `kubectl` 命令行工具比较特殊, 它同时支持 CA 双向认证与简单认证两种模式与 apiserver 通信, 其他客户端组件只能配置为双向安全认证或非安全模式与 apiserver 通信。

### 1) 基于 HTTP BASE 认证的配置过程

(1) 创建包括用户名、密码和 UID 的文件 `basic_auth_file`, 放置在合适的目录中, 例如 `/etc/kubernetes` 目录。需要注意的是, 这是一个纯文本文件, 用户名、密码都是明文。

```
# vi /etc/kubernetes/basic_auth_file
admin,admin,1
system,system,2
```

(2) 设置 kube-apiserver 的启动参数 “`--basic-auth-file`”, 使用上述文件提供安全认证:

```
--secure-port=6443
--basic-auth-file=/etc/kubernetes/basic_auth_file
```

然后, 重启 API Server 服务。

(3) 使用 `kubectl` 通过指定的用户名和密码来访问 API Server:

```
# kubectl --server=https://192.168.18.3:443 --username=admin --password=admin
--insecure-skip-tls-verify=true get nodes
```

## 2) 基于 TOKEN 认证的配置过程

(1) 创建包括用户名、密码和 UID 的文件 `token_auth_file`，放置在合适的目录中，例如 `/etc/kubernetes` 目录。需要注意的是，这是一个纯文本文件，用户名、密码都是明文。

```
$ cat /etc/kubernetes/token_auth_file
admin,admin,1
system,system,2
```

(2) 设置 kube-apiserver 的启动参数 “`--token-auth-file`”，使用上述文件提供安全认证：

```
--secure-port=6443
--token-auth-file=/etc/kubernetes/token_auth_file
```

然后，重启 API Server 服务。

(3) 用 curl 验证和访问 API Server：

```
$ curl -k --header "Authorization:Bearer admin" https://192.168.18.3:443/version
{
  "major": "1",
  "minor": "3",
  "gitVersion": "v1.3.3",
  "gitCommit": "c6411395e09da356c608896d3d9725acab821418",
  "gitTreeState": "clean",
  "buildDate": "2016-07-22T20:22:25Z",
  "goVersion": "go1.6.2",
  "compiler": "gc",
  "platform": "linux/amd64"
}
```

## 2.1.5 Kubernetes 集群的网络配置

在多个 Node 组成的 Kubernetes 集群内，跨主机的容器间网络互通是 Kubernetes 集群能够正常工作的前提条件。Kubernetes 本身并不会对跨主机的容器网络进行设置，这需要额外的工具来实现。除了谷歌公有云 GCE 平台提供的网络设置，一些开源的工具包括 Flannel、Open vSwitch、Weave、Calico 等都能够实现跨主机的容器间网络互通。随着 CNI 网络模型的逐渐成熟，Kubernetes 将优先使用 CNI 网络插件打通跨主机的容器网络。具体的网络原理和流行开源网络工具配置详见第 3 章的说明。

## 2.1.6 内网中的 Kubernetes 相关配置

Kubernetes 在能够访问 Internet 网络的环境中使用起来非常方便：一方面，在 `docker.io` 和 `gcr.io` 网站中已经存在了大量官方制作的 Docker 镜像；另一方面，GCE、AWS 提供的云平台已

经很成熟了，用户通过租用一定的空间来部署 Kubernetes 集群也很简便。

但是，许多企业内部由于安全性的原因无法访问 Internet。对于这些企业就需要通过创建一个内部的私有 Docker Registry，并修改一些 Kubernetes 的配置，来启动内网中的 Kubernetes 集群。

## 1. Docker Private Registry（私有 Docker 镜像库）

使用 Docker 提供的 Registry 镜像创建一个私有镜像仓库。

详细的安装步骤请参考 Docker 的官方文档 <https://docs.docker.com/registry/deploying/>。

## 2. kubelet 配置

由于在 Kubernetes 中是以 Pod 而不是以 Docker 容器为管理单元的，在 kubelet 创建 Pod 时，还通过启动一个名为 gcr.io/google\_containers/pause 的镜像来实现 Pod 的概念。

该镜像存在于谷歌镜像库 <http://gcr.io> 中，需要通过一台能够连上 Internet 的服务器将其下载，导出文件，再 push 到私有 Docker Registry 中。

之后，可以给每台 Node 的 kubelet 服务加上启动参数 `--pod-infra-container-image`，指定为私有 Docker Registry 中 pause 镜像的地址。例如：

```
# cat /etc/kubernetes/kubelet
KUBELET_ARGS="--api-servers=http://192.168.18.3:8080
--hostname-override=192.168.18.3 --log-dir=/var/log/kubernetes --v=2
--pod-infra-container-image=gcr.io/google_containers/pause-amd64:3.0"
```

如果该镜像无法从 gcr.io 下载，则也可以从 Docker Hub 上进行下载：

```
# docker pull kubeguide/pause-amd64:3.0
```

修改 kubelet 配置文件中的 `--pod_infra_container_image` 参数：

```
--pod-infra-container-image=kubeguide/pause-amd64:3.0
```

然后重启 kubelet 服务：

```
# systemctl restart kubelet
```

通过以上设置就在内网环境中搭建了一个企业内部的私有容器云平台。

### 2.1.7 Kubernetes 的版本升级

Kubernetes 的版本升级需要考虑到不要让当前集群中正在运行的容器受到影响。应对集群中的各 Node 逐个进行隔离，然后等待在其上运行的容器全部执行完成，再更新该 Node 上的 kubelet 和 kube-proxy 服务，将全部 Node 都更新完成后，最后更新 Master 的服务。

- 通过官网获取最新版本的二进制包 `kubernetes.tar.gz`，解压缩后提取服务的二进制文件。
- 逐个隔离 `Node`，等待在其上运行的全部容器工作完成后，更新 `kubelet` 和 `kube-proxy` 服务文件，然后重启这两个服务。
- 更新 `Master` 的 `kube-apiserver`、`kube-controller-manager`、`kube-scheduler` 服务文件并重启。

### 2.1.8 Kubernetes 核心服务配置详解

我们在 2.1.2 节对 Kubernetes 各服务启动进程的关键配置参数进行了简要说明，实际上 Kubernetes 的每个服务都提供了许多可配置的参数。这些参数涉及安全性、性能优化及功能扩展（Plugin）等方方面面。全面理解和掌握这些参数的含义和配置，无论对于 Kubernetes 的生产部署还是日常运维都有很大的帮助。

每个服务的可用参数都可以通过运行“`cmd --help`”命令进行查看，其中 `cmd` 为具体的服务启动命令，例如 `kube-apiserver`、`kube-controller-manager`、`kube-scheduler`、`kubelet`、`kube-proxy` 等。另外，可以通过在命令的配置文件（例如 `/etc/kubernetes/kubelet` 等）中添加“`--参数名=参数取值`”的语句来完成对某个参数的配置。

本节将对 Kubernetes 所有服务的参数进行全面介绍，为了方便学习和查阅，对每个服务的参数用一个小节进行详细说明。

#### 1. 公共配置参数

公共配置参数适用于所有服务，如表 2.3 所示的参数可用于 `kube-apiserver`、`kube-controller-manager`、`kube-scheduler`、`kubelet`、`kube-proxy`。本节对这些参数进行统一说明，不再在每个服务的参数列表中列出。

表 2.3 公共配置参数表

参数名和取值示例	说 明
<code>--log-backtrace-at traceLocation</code>	记录日志每到“file:行号”时打印一次 stack trace，默认值为 0
<code>--log-dir string</code>	日志文件路径
<code>--log-flush-frequency duration</code>	设置 flush 日志文件的时间间隔，默认值为 5s
<code>--logtostderr</code>	设置为 true 则表示将日志输出到 stderr，不输出到日志文件
<code>--alsologtostderr</code>	设置为 true 则表示将日志输出到文件的同时输出到 stderr
<code>--stderrthreshold severity</code>	将该 threshold 级别之上的日志输出到 stderr，默认值为 2
<code>--v Level</code>	glog 日志级别
<code>--vmodule moduleSpec</code>	glog 基于模块的详细日志级别，格式为 <code>pattern=N</code> ，以逗号分隔
<code>--version=[true]</code>	设置为 true，则将打印版本信息然后退出

## 2. kube-apiserver 启动参数

对 kube-apiserver 启动参数的详细说明如表 2.4 所示。

表 2.4 对 kube-apiserver 启动参数的详细说明

参数名和取值示例	说 明
--admission-control string	<p>对发送给 API Server 的任何请求进行准入控制，配置为一个“准入控制器”的列表，多个准入控制器时以逗号分隔。多个准入控制器将按顺序对发送给 API Server 的请求进行拦截和过滤，若某个准入控制器不允许该请求通过，则 API Server 拒绝此调用请求。可配置的准入控制器如下，默认值为 AlwaysAdmit。</p> <ul style="list-style-type: none"> <li>○ AlwaysAdmit: 允许所有请求。</li> <li>○ AlwaysDeny: 禁止所有请求，一般用于测试。</li> <li>○ AlwaysPullImages: 在启动容器之前总是去下载镜像，相当于在每个容器的配置项 imagePullPolicy=Always。</li> <li>○ DefaultStorageClass: 为了实现共享存储的动态供应，为未指定 StorageClass 或 PV 的 PVC 尝试匹配默认的 StorageClass。</li> <li>○ DefaultTolerationSeconds: 这个插件为那些没有设置 forgiveness tolerations 并具有 notready:NoExecute 和 unreachable:NoExecute 两种 taints 的 Pod 设置默认的“容忍”时间，为 5min。</li> <li>○ DenyEscalatingExec: 拦截所有 exec 和 attach 到具有特权的 Pod 上的请求。</li> <li>○ DenyExecOnPrivileged: 它会拦截所有想在 privileged container 上执行命令的请求。如果集群支持 privileged container，又希望限制用户在这些 privileged container 上执行命令，那么推荐使用该控制器。</li> <li>○ ImagePolicyWebhook: 这个插件将允许后端的一个 Webhook 程序来完成 admission controller 的功能。</li> <li>○ InitialResources: 实验性功能，用于当 Pod 未设置资源请求和资源限制时，根据该镜像历史资源使用数据自动为其设置资源请求。</li> <li>○ LimitPodHardAntiAffinityTopology: 用于 Pod 互斥性调度时，对 topologyKey 的限制，参见 2.3.9 节的说明。</li> <li>○ LimitRanger: 用于配额管理，作用于 Pod 与 Container 上，确保 Pod 与 Container 上的配额不会超标。</li> <li>○ NamespaceAutoProvision（已过时）：对所有请求校验 namespace，如果不存在，则自动创建该 namespace，推荐使用 NamespaceLifecycle。</li> <li>○ NamespaceExists（已过时）：对所有请求校验 namespace 是否已存在，如果不存在，则拒绝请求。已合并至 NamespaceLifecycle。</li> <li>○ NamespaceLifecycle: 如果尝试在一个不存在的 namespace 中创建资源对象，则该创建请求将被拒绝。删除一个 namespace 时，系统将会删除该 namespace 中的所有对象，包括 Pod、Service 等。</li> <li>○ PodPreset: 在 Pod 启动时注入应用所需的设置。</li> <li>○ PodSecurityPolicy: 在创建或修改 Pod 时决定是否根据 Pod 的 security context 和可用的 PodSecurityPolicy 对 Pod 的安全策略进行控制。</li> </ul>

续表

参数名和取值示例	说 明
	<ul style="list-style-type: none"> <li>◎ ResourceQuota: 用于配额管理, 作用于 Namespace 上, 它会监控所有的请求, 确保在 Namespace 上的配额不会超标。推荐在 admission control 参数列表中将这个插件排在最后一个。</li> <li>◎ SecurityContextDeny: 这个插件将使用了 SecurityContext 的 Pod 中定义的选项全部失效。SecurityContext 在 container 中定义了操作系统级别的安全设定 (uid、gid、capabilities、SELinux 等)。</li> <li>◎ ServiceAccount: 这个 plug-in 将 serviceAccounts 实现了自动化, 如果你想要使用 ServiceAccount 对象, 那么强烈推荐你使用它</li> </ul>
--admission-control="AlwaysAdmit"	<p>如果启用多种准入选项, 则建议加载的顺序如下所述。</p> <p>对 Kubernetes v1.6.0 及以上版本设置如下:</p> <pre>--admission-control=NamespaceLifecycle,LimitRanger,ServiceAccount,PersistentVolumeLabel,DefaultStorageClass,ResourceQuota,DefaultTolerationSeconds</pre> <p>对 Kubernetes &gt;= v1.4.0 版本:</p> <pre>--admission-control=NamespaceLifecycle,LimitRanger,ServiceAccount,DefaultStorageClass,ResourceQuota</pre> <p>对 Kubernetes &gt;= v1.2.0 版本:</p> <pre>--admission-control=NamespaceLifecycle,LimitRanger,ServiceAccount,ResourceQuota</pre> <p>对 Kubernetes &gt;= v1.0.0 版本:</p> <pre>--admission-control=NamespaceLifecycle,LimitRanger,SecurityContextDeny,ServiceAccount,PersistentVolumeLabel,ResourceQuota</pre>
--admission-control-config-file string	控制规则的配置文件
--advertise-address ip	用于广播给集群的所有成员自己的 IP 地址, 不指定该地址将使用 "--bind-address" 定义的 IP 地址
--allow-privileged	如果设置为 true, 则 Kubernetes 将允许在 Pod 中运行拥有系统特权的容器应用, 与 docker run --privileged 的效果相同
--anonymous-auth	设置为 true 时表示 APIServer 的安全端口可以接收匿名请求。不会被任何 authentication 拒绝的请求将被标记为匿名请求。匿名请求的用户名为 system:anonymous, 用户组为 system:unauthenticated。默认值为 true
--apiserver-count int	集群中运行的 API Server 数量, 默认值为 1
--audit-log-maxage int	审计日志文件保留最长天数
--audit-log-maxbackup int	审计日志文件个数
--audit-log-maxsize int	审计日志文件单个大小限制, 单位 MB, 默认为 100MB
--audit-log-path string	审计日志文件全路径
--authentication-token-webhook-cache-ttl duration	将 webhook token authenticator 返回的响应保存在缓存内的时间, 默认值为 2min (2m0s)
--authentication-token-webhook-config-file string	Webhook 相关的配置文件, 将用于 token authentication

续表

参数名和取值示例	说 明
--authorization-mode string	到 API Server 的安全访问的认证模式列表，以逗号分隔，可选值包括：AlwaysAllow、AlwaysDeny、ABAC、Webhook、RBAC，默认值为 AlwaysAllow
--authorization-policy-file string	当--authorization-mode 设置为 ABAC 时使用的 csv 格式的授权配置文件
--authorization-webhook-cache-authorized-ttl duration	将 webhook authorizer 返回的“已授权”响应保存在缓存内的时间，默认值为 5min (5m0s)
--authorization-webhook-cache-unauthorized-ttl duration	将 webhook authorizer 返回的“未授权”响应保存在缓存内的时间，默认值为 30s (30s)
--authorization-webhook-config-file string	当--authorization-mode 设置为 webhook 时使用的授权配置文件
--basic-auth-file string	设置该文件用于通过 HTTP 基本认证的方式访问 API Server 的安全端口
--bind-address ip	Kubernetes API Server 在本地地址的 6443 端口开启安全的 HTTPS 服务，默认值为 0.0.0.0
--cert-dir string	TLS 证书所在的目录，默认为/var/run/kubernetes。如果设置了--tls-cert-file 和 --tls-private-key-file，则该设置将被忽略，默认值为/var/run/kubernetes
--client-ca-file string	如果指定，则该客户端证书将被用于认证过程
--cloud-config string	云服务商的配置文件路径，不配置则表示不使用云服务商的配置文件
--cloud-provider string	云服务商的名称，不配置则表示不使用云服务商
--contention-profiling	
--cors-allowed-origins stringSlice	CORS（跨域资源共享）设置允许访问的源域列表，用逗号分隔，并可使用正则表达式匹配子网。如果不指定，则表示不启用 CORS
--default-not-ready-toleration-seconds int	等待 notReady:NoExecute 的 toleration 秒数，默认值为 300。默认会给所有未设置 toleration 的 Pod 上添加该设置。
--default-unreachable-toleration-seconds int	等待 unreachable:NoExecute 的 toleration 秒数，默认值为 300。默认会给所有未设置 toleration 的 Pod 上添加该设置。
--delete-collection-workers int	启动 DeleteCollection 的工作线程数，用于提高清理 namespace 的效率，默认值为 1
--deserialization-cache-size int	设置内存中缓存的 JSON 对象的个数
--enable-garbage-collector	设置为 true 表示启用垃圾回收器。必须与 kube-controller-manager 的该参数设置为相同的值，默认值为 true
--enable-swagger-ui	设置为 true 表示启用 swagger ui 网页，可通过 API Server 的 URL/swagger-ui 访问，默认值为 false
--etcd-cafile string	到 etcd 安全连接使用的 SSL CA 文件
--etcd-certfile string	到 etcd 安全连接使用的 SSL 证书文件
--etcd-keyfile string	到 etcd 安全连接使用的 SSL key 文件
--etcd-prefix string	在 etcd 中保存 Kubernetes 集群数据的根目录名，默认值为/registry
--etcd-quorum-read	设置为 true 表示启用 quorum read 机制
--etcd-servers stringSlice	以逗号分隔的 etcd 服务 URL 列表，etcd 服务以 http://ip:port 格式表示
--etcd-servers-overrides	按资源覆盖 etcd 服务的设置，以逗号分隔。单个覆盖格式为：group/resource #servers，其中 servers 格式为 http://ip:port，以分号分隔

续表

参数名和取值示例	说 明
--event-ttl duration	Kubernetes API Server 中各种事件（通常用于审计和追踪）在系统中保存的时间，默认为 1h（1h0m0s）
--experimental-bootstrap-token-auth	为实现 TLS bootstrapping 鉴权，在 'kube-system' namespace 中是否允许类型为 'bootstrap.kubernetes.io/token' 的 secrets。实验用
--experimental-keystone-ca-file string	设置使用 keystone 认证的证书路径，未设置则使用主机的 root CA 证书
--experimental-keystone-url string	设置 keystone 鉴权插件地址，实验用
--external-hostname string	用于生成该 Master 的对外 URL 地址，例如用于 swagger api 文档中的 URL 地址。
--feature-gates mapStringBool	用于实验性质的特性开关组，每个开关以 key=value 形式表示。当前可用的开关包括： Accelerators=true false (ALPHA - default=false) AffinityInAnnotations=true false (ALPHA - default=false) AllAlpha=true false (ALPHA - default=false) AllowExtTrafficLocalEndpoints=true false (BETA - default=true) AppArmor=true false (BETA - default=true) DynamicKubeletConfig=true false (ALPHA - default=false) DynamicVolumeProvisioning=true false (ALPHA - default=true) ExperimentalCriticalPodAnnotation=true false (ALPHA - default=false) ExperimentalHostUserNamespaceDefaulting=true false (BETA - default=false) StreamingProxyRedirects=true false (BETA - default=true) TaintBasedEvictions=true false (ALPHA - default=false)
--insecure-allow-any-token username/group1.group2	如果设置，则 server 将运行在非安全模式，允许接收任何 token，token 中的用户信息以 username/group1.group2 格式表示
--insecure-bind-address ip	绑定的不安全 IP 地址，与--insecure-port 共同使用，默认为 localhost。设置为 0.0.0.0 表示使用全部网络接口，默认值为 127.0.0.1
--insecure-port int	提供非安全认证访问的监听端口，默认值为 8080。应在防火墙中进行配置，以使得外部客户端不可以通过非安全端口访问 API Server
--ir-data-source string	设置 InitialResources 使用的数据源，可配置项包括 influxdb、gcm，默认值为 influxdb
--ir-dbname string	InitialResources 所需指标保存在 InfluxDB 中的数据库名称，默认值为 k8s
--ir-hawkular string	设置 Hawkular 的 URL 地址
--ir-influxdb-host string	InitialResources 所需指标所在 InfluxDB 的 URL 地址，默认值为 "localhost:8080/api/v1/namespaces/kube-system/services/monitoring-influxdb:api/proxy"
--ir-namespace-only	设置为 true 时表示基于相同 namespace 内的数据估算资源请求的值
--ir-password string	连接 InfluxDB 数据库的密码，默认值为 root
--ir-percentile int	InitialResources 进行资源估算时的采样百分比，实验用，默认值为 90
--ir-user string	连接 InfluxDB 数据库的用户名，默认值为 root
--kubelet-certificate-authority string	用于 CA 授权的 cert 文件路径
--kubelet-client-certificate string	用于 TLS 的客户端证书文件路径
--kubelet-client-key string	用于 TLS 的客户端 key 文件路径



续表

参数名和取值示例	说 明
--kubelet-https	指定 kubelet 是否使用 HTTPS 连接，默认值为 true
--kubelet-preferred-address-types stringSlice	连接 kubelet 时使用的节点地址类型 (NodeAddressTypes)，默认值为 [Hostname,InternalDNS,InternalIP,ExternalDNS,ExternalIP,LegacyHostIP]
--kubelet-read-only-port uint	已弃用，为 kubelet 端口号，默认值为 10255
--kubelet-timeout int	kubelet 执行操作的超时时间，默认值为 5s
--kubernetes-service-node-port int	设置 Master 服务是否使用 NodePort 模式，如果设置，则 Master 服务将映射到物理机的端口号；设置为 0 表示以 ClusterIP 的形式启动 Master 服务
--master-service-namespace string	已弃用，设置 Master 服务所在的 namespace，默认值为 default
--max-connection-bytes-per-sec int	设置为非 0 的值表示限制每个客户端连接的带宽为 xx 字节/s，目前仅用于需要长时间执行的请求
--max-mutating-requests-inflight int	同时处理的最大突变请求数量，默认值为 200，超过该数量的请求将被拒绝。设置为 0 表示无限制
--max-requests-inflight int	同时处理的最大请求数量，默认值为 400，超过该数量的请求将被拒绝。设置为 0 表示无限制
--min-request-timeout int	最小请求处理超时时间，单位为 s，默认值为 1800s，目前仅用于 watch request handler，其将会在该时间值上加一个随机时间作为请求的超时时间
--oidc-ca-file string	该文件内设置鉴权机构，OpenID Server 的证书将被其中一个机构进行验证。如果不设置，则将使用主机的 root CA 证书
--oidc-client-id string	OpenID Connect 的客户端 ID，在 oidc-issuer-url 设置时必须设置
--oidc-groups-claim string	定制的 OpenID Connect 用户组声明的设置，以字符串数组的形式表示，实验用
--oidc-issuer-url string	OpenID 发行者的 URL 地址，仅支持 HTTPS scheme，用于验证 OIDC JSON Web Token (JWT)
--oidc-username-claim string	OpenID claim 的用户名，默认值为 sub，实验用
--profiling	设置为 true 时表示打开性能分析，可以通过 <host>:<port>/debug/pprof/地址查看程序栈、线程等系统信息，默认值为 true
--repair-malformed-updates	设置为 true 表示服务器将尽可能修复无效或格式错误的 update request，以通过正确性校验，例如在一个 update request 中将一个已存在的 UID 值设置为空，默认值为 true
--requestheader-allowed-names stringSlice	允许的客户端证书中的 common names 列表，通过 header 中由"--requestheader-username-headers"参数指定的字段获取。未设置时表示由"--requestheader-client-ca-file"中认定的任意客户端证书都被允许
--requestheader-client-ca-file string	用于验证客户端证书的根证书，在信任"--requestheader-username-headers"参数中的用户名之前进行验证
--requestheader-extra-headers-prefix stringSlice	待审查请求 header 的前缀列表，建议用 X-Remote-Extra-
--requestheader-group-headers stringSlice	待审查请求 header 的用户组的列表，建议用 X-Remote-Group

续表

参数名和取值示例	说 明
--requestheader-username-headers stringSlice	待审查请求 header 的用户名的列表，通常用 X-Remote-User
--runtime-config mapStringString	一组 key=value 用于运行时的配置信息。apis/<groupVersion>/<resource> 可用于打开或关闭对某个 API 版本的支持。api/all 和 api/legacy 特别用于支持所有版本的 API 或支持旧版本的 API
--secure-port int	设置 API Server 使用的 HTTPS 安全模式端口号，设置为 0 表示不启用 HTTPS，默认值为 6443
--service-account-key-file stringArray	包含 PEM-encoded x509 RSA 公钥和私钥的文件路径，用于验证 Service Account 的 token。不指定则使用--tls-private-key-file 指定的文件
--service-account-lookup	设置为 true 时，系统会到 etcd 验证 ServiceAccount token 是否存在
--service-cluster-ip-range ipNet	Service 的 Cluster IP（虚拟 IP）池，例如 169.169.0.0/16，这个 IP 地址池不能与物理机所在的网络重合
--service-node-port-range portRange	Service 的 NodePort 能使用的主机端口号范围，默认值为 30000~32767，包括 30000 和 32767
--ssh-keyfile string	如果指定，则通过 SSH 使用指定的秘钥文件对 Node 进行访问
--ssh-user string	如果指定，则通过 SSH 使用指定的用户名对 Node 进行访问
--storage-backend string	设置持久化存储类型，可选项为 etcd2、etcd3，从 v1.6 版本开始默认值为 etcd3
--storage-media-type	持久化存储后端的介质类型。某些资源类型只能使用特定类型的介质进行保存，将忽略这个参数的设置，默认值为 application/vnd.kubernetes.protobuf
--storage-versions string	持久化存储的资源版本号，以分组形式标记，例如 "group1/version1,group2/version2,...", 默认值为 "apps/v1beta1,authentication.k8s.io/v1,authorization.k8s.io/v1,autoscaling/v1,batch/v1,certificates.k8s.io/v1beta1,componentconfig/v1alpha1,extensions/v1beta1,imagepolicy.k8s.io/v1alpha1,policy/v1beta1,rbac.authorization.k8s.io/v1beta1,settings.k8s.io/v1alpha1,storage.k8s.io/v1beta1,v1"
--target-ram-mb int	apiserver 的内存限制，单位为 MB，常用于设置缓存大小
--tls-cert-file string	包含 x509 证书的文件路径，用于 HTTPS 认证
--tls-private-key-file string	包含 x509 证书与 tls-cert-file 对应的私钥文件路径
--tls-sni-cert-key namedCertKey	x509 证书与私钥文件路径对，如果有多对设置，则需要指定多次--tls-sni-cert-key 参数，默认值为[]。可选配置域名后缀，例如"example.key.example.crt"或 "*.foo.com,foo.com:foo.key.foo.crt"
--token-auth-file string	用于访问 API Server 安全端口的 token 认证文件路径
--watch-cache	设置为 true 表示将 watch 进行缓存，默认值为 true
--watch-cache-sizes stringSlice	设置各资源对象 watch 缓存大小的列表，以逗号分隔，每个资源对象的设置格式为 resource#size，当 watch-cache 设置为 true 时生效

### 3. kube-controller-manager 启动参数

对 kube-controller-manager 启动参数的详细说明如表 2.5 所示。

表 2.5 对 kube-controller-manager 启动参数的详细说明

参数名和取值示例	说 明
--address=0.0.0.0	监听的主机 IP 地址，默认为 0.0.0.0 表示使用全部网络接口
--allocate-node-cidrs	设置为 true 表示使用云服务商为 Pod 分配的 CIDRs，仅用于公有云
--allow-verification-with-non-compliant-keys	设置为 true 时表示允许 SignatureVerifier 使用不符合 RFC6962 规范的密钥进行验证
--attach-detach-reconcile-sync-period duration	volume 的 attach/detach 等操作的 reconciler 同步等待时间，必须大于 1s，默认值为 1min（1m0s）
--cloud-config string	云服务商的配置文件路径，仅用于公有云
--cloud-provider string	云服务商的名称，仅用于公有云
--cluster-cidr string	集群中 Pod 的可用 CIDR 范围
--cluster-name string	集群的名称，默认值为 kubernetes
--cluster-signing-cert-file string	PEM-encoded X509 CA 证书文件，用于集群范围的认证，默认值为 "/etc/kubernetes/ca/ca.pem"
--cluster-signing-key-file string	PEM-encoded RSA 或 ECDSA 私钥文件，用于集群范围的认证，默认值为 "/etc/kubernetes/ca/ca.key"
--concurrent-deployment-syncs int32	设置允许的并发同步 deployment 对象的数量，值越大表示同步操作越快，但将会消耗更多的 CPU 和网络资源，默认值为 5
--concurrent-endpoint-syncs int32	设置并发执行 Endpoint 同步操作的数量，值越大表示同步操作越快，但将会消耗更多的 CPU 和网络资源，默认值为 5
--concurrent-gc-syncs int32	设置并发执行 GC Worker 的数量，默认值为 20
--concurrent-namespace-syncs int32	设置并发同步 namespace 对象的数量，值越大表示同步操作越快，但将会消耗更多的 CPU 和网络资源，默认值为 2
--concurrent-rc-syncs int32	并发执行 RC 同步操作的协程数，值越大表示同步操作越快，但将会消耗更多的 CPU 和网络资源，默认值为 5
--concurrent-replicaset-syncs int32	设置允许的并发同步 replica set 对象的数量，值越大表示同步操作越快，但将会消耗更多的 CPU 和网络资源，默认值为 5
--concurrent-resource-quota-syncs int32	设置允许的并发同步 replication controller 对象的数量，值越大表示同步操作越快，但将会消耗更多的 CPU 和网络资源
--concurrent-service-syncs int32	设置允许的并发同步 service 对象的数量，值越大表示同步操作越快，但将会消耗更多的 CPU 和网络资源，默认值为 1
--concurrent-serviceaccount-token-syncs int32	设置允许的并发同步 service account token 对象的数量，值越大表示同步操作越快，但将会消耗更多的 CPU 和网络资源，默认值为 1
--configure-cloud-routes	设置为 true 表示使用 allocate-node-cidrs 进行 CIDRs 的分配，仅用于公有云，默认值为 true

续表

参数名和取值示例	说 明
<code>--contention-profiling</code>	设置为 <code>true</code> 表示启用锁竞争性能数据采集，当 <code>profiling</code> 设置为 <code>true</code> 时生效
<code>--controller-start-interval duration</code>	启动各个 controller manager 的时间间隔，默认为 0s
<code>--controllers stringSlice</code>	要启用的 controller 列表，默认值为 <code>"*"</code> ，表示启用所有 controller， <code>'foo'</code> 表示启用名为 <code>'foo'</code> 的 controller， <code>'-foo'</code> 表示不启用名为 <code>'foo'</code> 的 controller。所有 controller 列表包括： <code>attachdetach</code> , <code>bootstrapsigner</code> , <code>certificatesigningrequests</code> , <code>cronjob</code> , <code>daemonset</code> , <code>deployment</code> , <code>disruption</code> , <code>endpoint</code> , <code>garbagecollector</code> , <code>horizontalpodautoscaling</code> , <code>job</code> , <code>namespace</code> , <code>node</code> , <code>persistentvolume-binder</code> , <code>podgc</code> , <code>replicaset</code> , <code>replicationcontroller</code> , <code>resourcequota</code> , <code>route</code> , <code>service</code> , <code>serviceaccount</code> , <code>serviceaccount-token</code> , <code>statefulset</code> , <code>tokencleaner</code> , <code>ttl</code> 。默认不启用的 controller 包括： <code>bootstrapsigner</code> , <code>tokencleaner</code>
<code>--deployment-controller-sync-period duration</code>	同步 deployment 的时间间隔，默认值为 30s
<code>--disable-attach-detach-reconcile-sync</code>	设置为 <code>true</code> 表示禁用 volume attach/detach 的同步操作，慎用
<code>--enable-dynamic-provisioning</code>	设置为 <code>true</code> 表示启用动态 provisioning（需底层存储驱动支持），默认值为 <code>true</code>
<code>--enable-garbage-collector</code>	设置为 <code>true</code> 表示启用垃圾回收机制，必须与 <code>kube-apiserver</code> 的该参数设置为相同的值，默认值为 <code>true</code>
<code>--enable-hostpath-provisioner</code>	设置为 <code>true</code> 表示启用 hostPath PV provisioning 机制，仅用于测试，不可用于多 Node 的集群环境
<code>--enable-taint-manager</code>	测试用，设置为 <code>true</code> 时表示启用 NoExecute Taints，并将在设置了该 taint 的 Node 上驱逐掉所有 not-tolerating 的 Pod，默认值为 <code>true</code>
<code>--feature-gates mapStringBool</code>	用于实验性质的特性开关组，每个开关以 <code>key=value</code> 形式表示。当前可用开关包括： <code>Accelerators=true/false</code> (ALPHA - default=false) <code>AffinityInAnnotations=true/false</code> (ALPHA - default=false) <code>AllAlpha=true/false</code> (ALPHA - default=false) <code>AllowExtTrafficLocalEndpoints=true/false</code> (BETA - default=true) <code>AppArmor=true/false</code> (BETA - default=true) <code>DynamicKubeletConfig=true/false</code> (ALPHA - default=false) <code>DynamicVolumeProvisioning=true/false</code> (ALPHA - default=true) <code>ExperimentalCriticalPodAnnotation=true/false</code> (ALPHA - default=false) <code>ExperimentalHostUserNamespaceDefaulting=true/false</code> (BETA - default=false) <code>StreamingProxyRedirects=true/false</code> (BETA - default=true) <code>TaintBasedEvictions=true/false</code> (ALPHA - default=false)
<code>--flex-volume-plugin-dir string</code>	设置 flex volume 插件应搜索其他第三方 volume 插件的全路径，默认值为 <code>"/usr/libexec/kubernetes/kubelet-plugins/volume/exec/"</code>
<code>--horizontal-pod-autoscaler-sync-period duration</code>	Pod 自动扩容器的 Pod 数量的同步时间间隔，默认值为 30s
<code>--horizontal-pod-autoscaler-use-rest-clients</code>	Alpha 版本的功能。设置为 <code>true</code> 时表示 HPA 将使用 kube-aggregator 提供的 REST 客户端，而不是通过 API Server proxy 进行访问。需要 HPA 能够支持自定义指标支持

续表

参数名和取值示例	说 明
--insecure-experimental-approve-all-kubelet-csrs-for-group string	Controller Manager 自动授权 kubelet 客户端证书 CSR 组
--kube-api-burst int32	发送到 API Server 的每秒的请求数量，默认值为 30
--kube-api-content-type string	发送到 API Server 的请求内容类型，默认值为 application/vnd.kubernetes.protobuf
--kube-api-qps float32	与 API Server 通信的 QPS 值，默认值为 20
--kubeconfig string	kubeconfig 配置文件路径，在配置文件中包括 Master 地址信息及必要的认证信息
--large-cluster-size-threshold int32	设置 Node 的数量，用于 NodeController 判定集群规模是否为大，用于 Pod Eviction 功能。设置该值后--secondary-node- eviction-rate 将会被隐式重置为 0。默认值为 50
--leader-elect	设置为 true 时表示进行 leader 选举，用于多个 Master 组件的高可用部署，默认值为 true
--leader-elect-lease-duration duration	leader 选举过程中非 leader 等待选举的时间间隔，默认值为 15s，当--leader-elect=true 时生效
--leader-elect-renew-deadline duration	leader 选举过程中在停止 leading 角色之前再次 renew 的时间间隔，应小于或等于 leader-elect-lease-duration，默认值为 10s，当--leader-elect=true 时生效
--leader-elect-retry-period duration	leader 选举过程中在获取 leader 角色和 renew 之间的等待时间，默认值为 2s，当--leader-elect=true 时生效
--master string	API Server 的 URL 地址，设置后不再使用 kubeconfig 中设置的值
--min-resync-period duration	最小重新同步的时间间隔，实际重新同步的时间为 MinResyncPeriod 到 2×MinResyncPeriod 之间的一个随机数，默认值为 12h0m0s
--namespace-sync-period duration	namespace 生命周期更新的同步时间间隔，默认值为 5m0s
--node-cidr-mask-size int32	Node CIDR 的子网掩码设置，默认值为 24
--node-eviction-rate float32	当 zone 仍为 healthy 状态（参考--unhealthy-zone-threshold 参数定义的健康状态，zone 指整个集群），在该 zone 中出现 Node 失效时，驱逐 Pod 时每秒处理的 Node 数量，默认值为 0.1
--node-monitor-grace-period duration	监控 Node 状态的时间间隔，默认值为 40s，超过该设置时间后，controller-manager 会把 Node 标记为不可用状态。此值的设置有如下要求： 它应该被设置为 kubelet 汇报的 Node 状态时间间隔（参数—node-status-update-frequency=10s）的 $N$ 倍， $N$ 为 kubelet 状态汇报的重试次数
--node-monitor-period duration	同步 NodeStatus 的时间间隔，默认值为 5s
--node-startup-grace-period duration	Node 启动的最大允许时间，超过此时间无响应则会标记 Node 为不可用状态（启动失败），默认值为 1m0s
--pod-eviction-timeout duration	在失效 Node 上删除 Pod 的超时时间，默认值为 5m0s
--port int32	controller-manager 监听的主机端口号，默认值为 10252
--profiling	打开性能分析，可以通过<host>:<port>/debug/pprof/地址查看程序栈、线程等系统运行信息
--pv-recycler-increment-timeout-nfs int32	使用 nfs scrubber 的 Pod 每增加 1Gi 空间在 ActiveDeadlineSeconds 上增加的时间，默认值为 30s

续表

参数名和取值示例	说 明
--pv-recycler-minimum-timeout-hostpath int32	使用 hostPath recycler 的 Pod 的最小 ActiveDeadlineSeconds 秒数，默认值为 60s。实验用
--pv-recycler-minimum-timeout-nfs int32	使用 nfs recycler 的 Pod 的最小 ActiveDeadlineSeconds 秒数，默认值为 300s
--pv-recycler-pod-template-filepath-host path string	使用 hostPath recycler 的 Pod 的模板文件全路径
--pv-recycler-pod-template-filepath-nfs string	使用 nfs recycler 的 Pod 的模板文件全路径
--pv-recycler-timeout-increment-hostpath int32	使用 hostPath scrubber 的 Pod 每增加 1Gi 空间在 ActiveDeadlineSeconds 上增加的时间，默认值为 30s
--pvclaimbinder-sync-period duration	同步 PV 和 PVC（容器声明的 PV）的时间间隔，默认值为 15s
--resource-quota-sync-period duration	resource quota 使用信息同步的时间间隔，默认值为 5m0s
--root-ca-file string	根 CA 证书文件路径，将被用于 Service Account 的 token secret 中
--route-reconciliation-period duration	云服务商创建的路由同步时间，默认值为 10s
--secondary-node-eviction-rate float32	当 zone 为 unhealthy 状态（参考--unhealthy-zone-threshold 参数定义的健康状态，zone 指整个集群），在该 zone 中出现 Node 失效时，驱逐 Pod 时每秒处理的 Node 数量，默认值为 0.01。当设置了--large-cluster-size-threshold 参数并且集群 Node 数量少于--large-cluster-size-threshold 的值时，该参数被隐式重置为 0
--service-account-private-key-file string	用于为 Service Account token 签名的 PEM-encoded RSA 私钥文件路径
--service-cluster-ip-range string	Service 的 IP 范围
--service-sync-period string	同步 service 与外部 load balancer 的时间间隔，默认值为 5m0s
--terminated-pod-gc-threshold int32	设置可保存的终止 Pod 的数量，超过该数量时，垃圾回收器将开始进行删除操作。设置为不大于 0 的值表示不启用该功能，默认值为 12500
--unhealthy-zone-threshold float32	设置一个 zone 中多少比例的 Node 失效将被判断为 unhealthy，至少有 3 台 Node 失效才能进行判断，默认值为 0.55
--use-service-account-credentials	设置为 true 时表示为每个 controller 分别设置 service account

4. kube-scheduler 启动参数

对 kube-scheduler 启动参数的详细说明如表 2.6 所示。

表 2.6 对 kube-scheduler 启动参数的详细说明

参数名和取值示例	说 明
--address string	监听的主机 IP 地址，默认值为 0.0.0.0 表示使用全部网络接口
--algorithm-provider string	设置调度算法，可选项为 ClusterAutoscalerProvider 或 DefaultProvider，默认值为 DefaultProvider
--contention-profiling	设置为 true 表示启用锁竞争性性能数据采集，当 profiling 被设置为 true 时生效

续表

参数名和取值示例	说 明
--feature-gates mapStringBool	用于实验性质的特性开关组，每个开关以 key=value 形式表示。当前可用开关包括： Accelerators=true false (ALPHA - default=false) AffinityInAnnotations=true false (ALPHA - default=false) AllAlpha=true false (ALPHA - default=false) AllowExtTrafficLocalEndpoints=true false (BETA - default=true) AppArmor=true false (BETA - default=true) DynamicKubeletConfig=true false (ALPHA - default=false) DynamicVolumeProvisioning=true false (ALPHA - default=true) ExperimentalCriticalPodAnnotation=true false (ALPHA - default=false) ExperimentalHostUserNamespaceDefaulting=true false (BETA - default=false) StreamingProxyRedirects=true false (BETA - default=true) TaintBasedEvictions=true false (ALPHA - default=false)
--hard-pod-affinity-symmetric-weight int	表示 Pod 调度规则亲和力的权重值，取值范围为 0~100。RequiredDuringScheduling 亲和性是非对称的，但对每一个 RequiredDuringScheduling 亲和性都存在一个对应的隐式 PreferredDuringScheduling 亲和性规则。该设置表示隐式 PreferredDuringScheduling 亲和性规则的权重值，默认值为 1
--kube-api-burst int32	发送到 API Server 的每秒请求数量，默认值为 100
--kube-api-content-type string	发送到 API Server 的请求内容类型，默认值为 application/vnd.kubernetes.protobuf
--kube-api-qps float32	与 API Server 通信的 QPS 值，默认值为 50
--kubeconfig string	kubeconfig 配置文件路径，在配置文件中包括 Master 的地址信息及必要的认证信息
--leader-elect	设置为 true 表示进行 leader 选举，用于多个 Master 组件的高可用部署，默认值为 true
--leader-elect-lease-duration duration	leader 选举过程中非 leader 等待选举的时间间隔，默认值为 15s，当 leader-elect=true 时生效
--leader-elect-renew-deadline duration	leader 选举过程中在停止 leading 角色之前再次 renew 的时间间隔，应小于或等于 leader-elect-lease-duration，默认值为 10s，当 leader-elect=true 时生效
--leader-elect-retry-period duration	leader 选举过程中获取 leader 角色和 renew 之间的等待时间，默认值为 2s，当 leader-elect=true 时生效
--master string	API Server 的 URL 地址，设置后不再使用 kubeconfig 中设置的值
--policy-config-file string	调度策略（scheduler policy）配置文件的路径
--port int32	scheduler 监听的主机端口号，默认值为 10251
--profiling	打开性能分析，可以通过 <host>:<port>/debug/pprof/地址查看栈、线程等系统运行信息，默认值为 true
--scheduler-name string	调度器名称，用于选择哪些 Pod 将被该调度器进行处理，选择的依据是 Pod 的 annotation 设置，包含 key='scheduler.alpha.kubernetes.io/name' 的 annotation，默认值为 default-scheduler

5. kubelet 启动参数

对 kubelet 启动参数的详细说明如表 2.7 所示。

表 2.7 对 kubelet 启动参数的详细说明

参数名和取值示例	说 明
--address ip	绑定主机 IP 地址，默认值为 0.0.0.0 表示使用全部网络接口
--allow-privileged	是否允许以特权模式启动容器，默认值为 false
--api-servers	API Server 地址列表，以 ip:port 格式表示，以逗号分隔
--anonymous-auth	设置为 true 时表示 Kubelet Server 可以接收匿名请求。不会被任何 authentication 拒绝的请求将被标记为匿名请求。匿名请求的用户名为 system:anonymous，用户组为 system:unauthenticated。默认值为 true
--application-metrics-count-limit int	为每个容器保存的性能指标的最大数量，默认值为 100
--authentication-token-webhook	使用 TokenReview API 授权客户端 token
--authentication-token-webhook-cache-ttl duration	将 webhook token authenticator 返回的响应保存在缓存内的时间，默认值为 2min (2m0s)
--authorization-mode string	到 Kubelet Server 的安全访问的认证模式，可选值包括：AlwaysAllow、Webhook（使用 SubjectAccessReview API 进行授权），默认值为 AlwaysAllow
--authorization-webhook-cache-authorized-ttl duration	webhook authorizer 返回“已授权”的应答缓存时间，默认值为 5m0s
--authorization-webhook-cache-unauthorized-ttl duration	webhook authorizer 返回“未授权”的应答缓存时间，默认值为 30s
--azure-container-registry-config string	Azure 云上镜像库的配置文件路径
--boot-id-file string	以逗号分隔的文件列表，使用第 1 个存在 boot-id 的文件，默认值为 /proc/sys/kernel/random/boot_id
--cadvisor-port int32	本地 cAdvisor 监听的端口号，默认值为 4194
--cert-dir string	TLS 证书所在的目录，默认值为 /var/run/kubernetes。如果设置了 --tls-cert-file 和 --tls-private-key-file，则该设置将被忽略
--cgroup-driver string	用于操作本机 cgroup 的驱动模式，支持的选项包括 groupfs 或 systemd，默认值为 cgroupfs
--cgroup-root string	为 pods 设置的 root cgroup，如果不设置，则将使用容器运行时的默认设置，默认值为空字符串（表示为两个单引号"）
--cgroups-per-qos	设置为 true 表示启用创建 QoS cgroup hierarchy，默认值为 true
--chaos-chance float	随机产生客户端错误的概率，用于测试目的，默认值为 0.0，即不产生
--client-ca-file	设置客户端 CA 证书文件，一旦设置该文件，则将对所有客户端请求进行鉴权，验证客户端证书的 CommonName 信息
--cloud-config string	云服务商的配置文件路径
--cloud-provider string	云服务商的名称，默认将自动检测，设置为空表示无云服务商，默认值为 auto-detect



续表

参数名和取值示例	说 明
--cluster-dns stringSlice	集群内 DNS 服务的 IP 地址,以逗号分隔。仅当 Pod 设置了“dnsPolicy=ClusterFirst”属性时可用。注意,所有 DNS 服务器必须包含相同的记录组,否则名字解析可能会出错
--cluster-domain string	集群内 DNS 服务所用域名
--cni-bin-dir string	Alpha 版特性,CNI 插件二进制文件所在的目录,默认值为/opt/cni/bin
--cni-conf-dir string	Alpha 版特性,CNI 插件配置文件所在的目录,默认值为/etc/cni/net.d
--container-hints	容器 hints 文件所在的全路径,默认值为/etc/cadvisor/container_hints.json
--container-runtime string	容器类型,目前支持 Docker、rkt,默认值为 docker
--container-runtime-endpoint string	实验性特性,容器运行时的远程 unix socket endpoint,仅当启用 CRI 时生效(--enable-cri=true)
--containerized	将 kubelet 运行在容器中,仅供测试使用,默认值为 false
--contention-profiling	
--cpu-cfs-quota	设置为 true 表示启用 CPU CFS quota,用于设置容器的 CPU 限制,默认值为 true
--docker string	Docker 服务的 Endpoint 地址,默认值为 unix:///var/run/docker.sock
--docker-endpoint string	Docker 服务的 Endpoint 地址,默认值为 unix:///var/run/docker.sock
--docker-env-metadata-whitelist string	Docker 容器需要使用的环境变量 key 列表,以逗号分隔
--docker-only	设置为 true,表示仅报告 Docker 容器的统计信息而不再报告其他统计信息
--docker-root string	已弃用,Docker 根目录的全路径,默认值为/var/lib/docker
--enable-controller-attach-detach	设置为 true 表示启用 Attach/Detach Controller 进行调度到该 Node 的 volume 的 attach 与 detach 操作,同时禁用 kubelet 执行 attach、detach 的操作,默认值为 true
--enable-custom-metrics	设置为 true 表示启用采集自定义性能指标
--enable-debugging-handlers	设置为 true 表示提供远程访问本节点容器的日志、进入容器执行命令等相关 Rest 服务,默认值为 true
--enable-load-reader	设置为 true 表示启用 CPU 负载的 reader
--enable-server	启动 kubelet 上的 http rest server,此 server 提供了获取本节点上运行的 Pod 列表、Pod 状态和其他管理监控相关的 Rest 接口
--enforce-node-allocatable stringSlice	本 Node 上 kubelet 资源的分配设置,以逗号分隔,可选配置为'pods'、'system-reserved'和'kube-reserved'。设置了'system-reserved'和'kube-reserved'这两个值时,要求同时设置'--system-reserved-cgroup'和'--kube-reserved-cgroup'这两个参数。参考 <a href="https://github.com/kubernetes/community/blob/master/contributors/design-proposals/node-allocatable.md">https://github.com/kubernetes/community/blob/master/contributors/design-proposals/node-allocatable.md</a> 。默认值为'pods'
--event-burst int32	临时允许的 Event 记录突发的最大数量,默认值为 10,当 event-qps>0 时生效
--event-qps int32	设置大于 0 的值表示限制每秒能创建出的 Event 数量,设置为 0 表示不限制,默认值为 5
--event-storage-age-limit string	保存 Event 的最大时间。按事件类型以 key=value 的格式表示,以逗号分隔,事件类型包括 creation、oom 等,“default”表示所有事件的类型,默认值为“default=0”

续表

参数名和取值示例	说 明
--event-storage-event-limit string	保存 Event 的最大数量。按事件类型以 key=value 格式表示，以逗号分隔，事件类型包括 creation、oom 等，“default”表示所有事件的类型，默认值为"default=0"
--eviction-hard string	触发 Pod Eviction 操作的一组硬门限设置，默认值为"memory.available<100Mi"
--eviction-max-pod-grace-period int32	终止 Pod 操作为 Pod 自行停止预留的时间，单位为 s。时间到达时，将触发 Pod Eviction 操作。默认值为 0，设置为负数表示使用 Pod 中指定的值
--eviction-minimum-reclaim string	当本节点压力过大时，kubelet 进行 Pod Eviction 操作，进而需要完成资源回收的最小数量的一组设置，例如 imagefs.available=2Gi
--eviction-pressure-transition-period duration	kubelet 在触发 Pod Eviction 操作之前等待的最长时间，默认值为 5m0s
--eviction-soft string	触发 Pod Eviction 操作的一组软门限设置，例如可用内存<1.5Gi，与 grace-period 一起生效，当 Pod 的响应时间超过 grace-period 后进行触发
--eviction-soft-grace-period string	触发 Pod Eviction 操作的一组软门限等待时间设置，例如 memory.available=1m30s
--exit-on-lock-contention	设置为 true 表示当有文件锁存在时 kubelet 也可以退出
--experimental-allocatable-ignore-eviction	设置为 true 表示计算 Node Allocatable 时忽略硬门限设置。参考 <a href="https://github.com/kubernetes/community/blob/master/contributors/design-proposals/node-allocatable.md">https://github.com/kubernetes/community/blob/master/contributors/design-proposals/node-allocatable.md</a> 。默认值为 false
--experimental-allowed-unsafe-sysctls stringSlice	[实验性特性] 不安全 sysctl 或不安全 sysctl 模式（以*结尾）白名单，以逗号分隔
--experimental-bootstrap-kubeconfig string	[实验性特性] kubelet 获取客户端证书的 kubeconfig 文件。如果--kubeconfig 指定的文件不存在，将使用本参数指定的配置到 API Server 获取客户端证书。成功获取后在--kubeconfig 路径生成 kubeconfig 文件。证书和密钥文件将被保存到 --cert-dir 指定的目录下
--experimental-check-node-capabilities-before-mount	[实验性特性] 设置为 true 表示 kubelet 在进行 mount 操作之前对本 Node 上所需组件（二进制文件等）进行检查
--experimental-fail-swap-on	[实验性特性] 设置为 true 表示在 swap 设置为 on 的 node 上不启动 kubelet
--experimental-kernel-memcg-notification	[实验性特性] 设置为 true 表示 kubelet 将会集成 kernel 的 memcg 通知机制，以判断达到了内存 Eviction 门限
--experimental-mounter-path string	[实验性特性] mounter 二进制文件的路径。设置为空表示使用默认 mount
--experimental-qos-reserved mapStringString	[实验性特性] 一组 ResourceName=Percentage（例如 memory=50%）的设置用于描述在某个 QoS 等级对 pod 资源请求的预留设置。当前版本仅支持 memory 的设置。默认值为 none
--feature-gates string	用于实验性质的特性开关组，每个开关以 key=value 形式表示。当前可用开关包括： Accelerators=true/false (ALPHA - default=false) AffinityInAnnotations=true/false (ALPHA - default=false) AllAlpha=true/false (ALPHA - default=false) AllowExtTrafficLocalEndpoints=true/false (BETA - default=true) AppArmor=true/false (BETA - default=true)

续表

参数名和取值示例	说 明
	DynamicKubeletConfig=true/false (ALPHA - default=false) DynamicVolumeProvisioning=true/false (ALPHA - default=true) ExperimentalCriticalPodAnnotation=true/false (ALPHA - default=false) ExperimentalHostUserNamespaceDefaulting=true/false (BETA - default=false) StreamingProxyRedirects=true/false (BETA - default=true) TaintBasedEvictions=true/false (ALPHA - default=false)
--file-check-frequency duration	在 File Source 作为 Pod 源的情况下, kubelet 定期重新检查文件变化的时间间隔, 文件发生变化后, kubelet 重新加载更新的文件内容, 默认值为 20s
--global-housekeeping-interval duration	全局 housekeeping 的时间间隔, 默认值为 1m0s
--google-json-key string	用于谷歌的云平台 Service Account 进行用于签名的 JSON key
--hairpin-mode string	设置 hairpin 模式, 表示 kubelet 设置 hairpin NAT 的方式。该模式允许后端 Endpoint 在访问其本身 Service 时能够再次 loadbalance 回自身。可选项包括 promiscuous-bridge、hairpin-veth 和 none, 默认值为"promiscuous-bridge"
--healthz-bind-address ip	healthz 服务监听的 IP 地址, 默认值为 127.0.0.1, 设置为 0.0.0.0 表示监听全部 IP 地址
--healthz-port int32	本地 healthz 服务监听的端口号, 默认值为 10248
--host-ipc-sources stringSlice	kubelet 允许 Pod 使用宿主机 ipc namespace 的列表, 以逗号分隔, 默认值为[*]
--host-network-sources stringSlice	kubelet 允许 Pod 使用宿主机 network 的列表, 以逗号分隔, 默认值为[*]
--host-pid-sources stringSlice	kubelet 允许 Pod 使用宿主机 pid namespace 的列表, 以逗号分隔, 默认值为[*]
--hostname-override string	设置本 Node 在集群中的主机名, 不设置将使用本机 hostname
--housekeeping-interval duration	对容器进行 housekeeping 操作的时间间隔, 默认值为 10s
--http-check-frequency duration	HTTP URL Source 作为 Pod 源的情况下, kubelet 定期检查 URL 返回的内容是否发生变化的时间周期, 作用同 file-check-frequency 参数, 默认值为 20s
--image-gc-high-threshold int32	镜像垃圾回收上限, 磁盘使用空间达到该百分比时, 镜像垃圾回收将持续工作, 默认值为 90
--image-gc-low-threshold int32	镜像垃圾回收下限, 磁盘使用空间在达到该百分比之前, 镜像垃圾回收将不启用, 80
--image-pull-progress-deadline duration	如果在该参数值之前还未能开始 pull 镜像的过程, pull 镜像操作将被取消, 默认值为 1m0s
--image-service-endpoint string	[实验性特性] 远程镜像服务的 unix socket endpoint。未设定则使用--container-runtime-endpoint 的值。仅当启用 CRI 集成 (--enable-cri=true) 时生效
--iptables-drop-bit int32	标记数据包将被丢弃 (Drop) 的 fwmark 位设置, 有效范围为[0, 31], 默认值为 15
--iptables-masquerade-bit int32	标记数据包将进行 SNAT 的 fwmark 位设置, 有效范围为[0, 31], 必须与 kube-proxy 的相关参数设置一致, 默认值为 14
--keep-terminated-pod-volumes	设置为 true 表示 pod 被删除后仍然保留之前 mount 过的 volumes, 常用于 volume 相关问题的查错
--kube-api-burst int32	发送到 API Server 的每秒请求数量, 默认值为 10

续表

参数名和取值示例	说 明
--kube-api-content-type string	发送到 API Server 的请求内容类型, 默认值为"application/vnd.kubernetes.protobuf"
--kube-api-qps int32	与 API Server 通信的 QPS 值, 默认值为 5
--kube-reserved mapStringString	kubernetes 系统预留的资源配置, 以一组 ResourceName=ResourceQuantity 格式表示, 例如 cpu=200m,memory=150G。目前仅支持 CPU 和内存的设置, 详见 <a href="http://releases.k8s.io/HEAD/docs/user-guide/compute-resources.md">http://releases.k8s.io/HEAD/docs/user-guide/compute-resources.md</a> , 默认值为 none
--kube-reserved-cgroup string	用于管理 kubernetes 的带 '--kube-reserved' 标签组件的计算资源, 设置顶层 cgroup 全路径名, 例如 '/kube-reserved', 默认值为 "
--kubeconfig string	kubeconfig 配置文件路径, 在配置文件中包括 Master 地址信息及必要的认证信息, 默认值为 "/var/lib/kubelet/kubeconfig"
--kubelet-cgroups string	kubelet 运行所需的 cgroups 名称
--lock-file string	[ALPHA 版特性] kubelet 使用的 lock 文件
--log-cadvisor-usage	设置为 true 表示将 cAdvisor 容器的使用情况进行日志记录
--machine-id-file string	用于查找 machine-id 的文件列表, 使用找到的第 1 个值, 默认值为 "/etc/machine-id,/var/lib/dbus/machine-id"
--make-iptables-util-chains	设置为 true 表示 kubelet 将确保 Iptables 规则在 node 上存在, 默认值为 true
--manifest-url string	为 HTTP URL Source 源类型时, kubelet 用来获取 Pod 定义的 URL 地址, 此 URL 返回一组 Pod 定义
--manifest-url-header string	访问 manifest URL 地址时使用的 HTTP 头信息, 以 key:value 格式表示
--master-service-namespace string	Master 服务的命名空间, 默认值为 "default"
--max-open-files int	kubelet 打开的最大文件数量, 默认值为 1000 000
--max-pods int32	kubelet 能运行的最大 Pod 数量, 默认值为 110
--minimum-image-ttl-duration duration	不再使用的镜像在被清理之前的最小存活时间, 例如 300ms、10s 或 2h45m, 超过此存活时间的镜像被标记为可被 GC 清理, 默认值为 2m0s
--network-plugin string	[Alpha 版特性] 自定义的网络插件的名字, Pod 的生命周期中相关的一些事件会调用此网络插件进行处理
--network-plugin-dir string	[Alpha 版特性] 扫描网络插件的目录
--network-plugin-mtu int32	[Alpha 版特性] 传递给网络插件的 MTU 值, 设置为 0 表示使用默认 1460 MTU
--node-ip string	设置本 Node 的 IP 地址
--node-labels mapStringString	[Alpha 版特性] kubelet 注册本 Node 时设置的 Labels, label 以 key=value 的格式表示, 多个 label 以逗号分隔
--node-status-update-frequency duration	kubelet 向 Master 汇报 Node 状态的时间间隔, 默认值为 10s。与 controller-manager 的 --node-monitor-grace-period 参数共同起作用
--non-masquerade-cidr string	kubelet 向该 IP 段之外的 IP 地址发送的流量将使用 IP Masquerade 技术, 默认值为 "10.0.0.0/8"
--oom-score-adj int32	kubelet 进程的 oom_score_adj 参数值, 有效范围为 [-1000, 1000], 默认值为 -999

续表

参数名和取值示例	说 明
--pod-cidr string	用于给 Pod 分配 IP 地址的 CIDR 地址池，仅在单机模式中使用。在一个集群中，kubelet 会从 API Server 中获取 CIDR 设置
--pod-infra-container-image string	用于 Pod 内网络命名空间共享的基础 pause 镜像，默认值为"gcr.io/google_containers/pause-amd64:3.0"
--pod-manifest-path string	Pod Manifest 文件路径，不扫描以'.'开头的隐藏文件
--pods-per-core int32	该 kubelet 上每个 core 可运行的 Pod 数量。最大值将被 max-pods 参数限制。默认值为 0 表示不做限制
--port int32	kubelet 服务监听的本机端口号，默认值为 10250
--protect-kernel-defaults	设置 kernel tuning 的默认 kubelet 行为。如果 kernel tunables 与 kubelet 默认值不同，kubelet 将报错。
--read-only-port int32	kubelet 服务监听的“只读”端口号，默认值为 10255，设置为 0 表示不启用
--really-crash-for-testing	设置为 true 表示发生 panics 情况时崩溃，仅用于测试
--register-node	将本 Node 注册到 API Server，默认值为 true
--register-with-taints []api.Taint	设置本 Node 的 taints，格式为"<key>=<value>:<effect>"，以逗号分隔。当 --register-node=false 时不生效
--registry-burst int32	最多同时拉取镜像的数量，默认值为 10
--registry-qps int32	在 Pod 创建过程中容器的镜像可能需要从 Registry 中拉取，由于拉取镜像的过程中会消耗大量带宽，因此可能需要限速，此参数与 registry-burst 一起用来限制每秒拉取多少个镜像，默认值为 5
--require-kubeconfig	设置为 true 表示如果 config 中有错，kubelet 将推出，同时忽略--api-servers 的设置
--resolv-conf string	命名服务配置文件，用于容器内应用的 DNS 解析，默认值为"/etc/resolv.conf"
--rkt-api-endpoint string	rkt API 服务的 URL 地址，--container-runtime='rkt' 时生效，默认值为 "localhost:15441"
--rkt-path string	rkt 二进制文件的路径，不指定的话从环境变量 \$PATH 中查找，--container-runtime='rkt' 时生效
--root-dir string	kubelet 运行根目录，将保持 Pod 和 volume 的相关文件，默认值为/var/lib/kubelet"
--runonce	设置为 true 表示创建完 Pod 之后立即退出 kubelet 进程，与--api-servers 和 --enable-server 参数互斥
--runtime-cgroups string	为容器 runtime 设置的 cgroup
--runtime-request-timeout duration	除了长时间运行的 request，对其他 request 的超时时间设置，包括 pull、logs、exec、attach 等操作。当超时时间到达时，请求会被杀掉，抛出一个错误并会重试。默认值为 2m0s
--seccomp-profile-root string	seccomp 配置文件目录，默认值为"/var/lib/kubelet/seccomp"
--serialize-image-pulls	按顺序挨个 pull 镜像。建议 Docker 低于 <v1.9 版本或使用 aufs storage backend 时设置为 true，详见 issue #10959，默认值为 true
--storage-driver-buffer-duration duration	将缓存数据写入后端存储的时间间隔，默认值为 1m0s

续表

参数名和取值示例	说 明
--storage-driver-db string	后端存储的数据库名称，默认值为"cadvisor"
--storage-driver-host string	后端存储的数据库连接 URL 地址，默认值为"localhost:8086"
--storage-driver-password string	后端存储的数据库密码，默认值为"root"
--storage-driver-secure	后端存储的数据库是否用安全连接，默认值为 false
--storage-driver-table string	后端存储的数据库表名，默认值为"stats"
--storage-driver-user string	后端存储的数据库用户名，默认值为"root"
--streaming-connection-idle-timeout duration	在容器中执行命令或者进行端口转发的过程中会产生输入、输出流，这个参数用来控制连接空闲超时而关闭的时间，如果设置为 5m，则表示连接超过 5min 没有输入、输出的情况下就被认为是空闲的，而会被自动关闭。默认值为 4h0m0s
--sync-frequency duration	同步运行中容器的配置的频率，默认值为 1m0s
--system-cgroups /	kubelet 为运行非 kernel 进程设置的 cgroups 名称，默认值为""
--system-reserved mapStringString	系统预留的资源配置，以一组 ResourceName=ResourceQuantity 格式表示，例如 cpu=200m,memory=500M。目前仅支持 CPU 和内存的设置，详见 <a href="http://releases.k8s.io/HEAD/docs/user-guide/compute-resources.md">http://releases.k8s.io/HEAD/docs/user-guide/compute-resources.md</a> ，默认值为 none
--system-reserved-cgroup string	用于管理非 kubernetes 的带"--system-reserved"标签组件的计算资源，设置顶层 cgroup 全路径名，例如'/system-reserved'，默认值为"
--tls-cert-file string	包含 x509 证书的文件路径，用于 HTTPS 认证
--tls-private-key-file string	包含 x509 与 tls-cert-file 对应的私钥文件路径
--volume-plugin-dir string	[Alpha 版特性]搜索第三方 volume 插件的目录，默认值为"/usr/libexec/kubernetes/kubelet-plugins/volume/exec/"
--volume-stats-aggr-period duration	kubelet 计算所有 Pod 和 volume 的磁盘使用情况聚合值的时间间隔，默认值为 1m0s。设置为 0 表示不启用该计算功能

6. kube-proxy 启动参数

kube-proxy 启动参数的详细说明见表 2.8。

表 2.8 对 kube-proxy 启动参数的详细说明

参数名和取值示例	说 明
--bind-address ip	kube-proxy 绑定主机的 IP 地址，默认值为 0.0.0.0 表示绑定所有 IP 地址
--cleanup-iptables	设置为 true 表示清除 Iptables 规则后退出
--cluster-cidr string	集群中 Pod 的 CIDR 地址范围，用于桥接集群外部流量到内部。用于公有云环境
--config-sync-period duration	从 API Server 更新配置的时间间隔，必须大于 0，默认值为 15m0s
--conntrack-max-per-core=32768	跟踪每个 CPU core 的 NAT 连接的最大数量（设置为 0 时表示无限制，并忽略 conntrack-min 的值），默认值为 32768
--conntrack-min int32	最小 conntrack 条目的分配数量，默认值为 131072
--conntrack-tcp-timeout-close-wait duration	当 TCP 连接处于 CLOSE_WAIT 状态时的 NAT 超时时间，默认值为 1h0m0s

续表

参数名和取值示例	说 明
--conntrack-tcp-timeout-established	建立 TCP 连接的超时时间，设置为 0 表示无限制，默认值为 24h0m0s
--feature-gates mapStringBool	用于实验性质的特性开关组，每个开关以 key=value 形式表示。当前可用开关包括： Accelerators=true false (ALPHA - default=false) AffinityInAnnotations=true false (ALPHA - default=false) AllAlpha=true false (ALPHA - default=false) AllowExtTrafficLocalEndpoints=true false (BETA - default=true) AppArmor=true false (BETA - default=true) DynamicKubeletConfig=true false (ALPHA - default=false) DynamicVolumeProvisioning=true false (ALPHA - default=true) ExperimentalCriticalPodAnnotation=true false (ALPHA - default=false) ExperimentalHostUserNamespaceDefaulting=true false (BETA - default=false) StreamingProxyRedirects=true false (BETA - default=true) TaintBasedEvictions=true false (ALPHA - default=false)
--healthz-bind-address ip	healthz 服务绑定主机 IP 地址，默认值为 127.0.0.1，设置为 0.0.0.0 表示使用所有 IP 地址
--healthz-port int32	healthz 服务监听的主机端口号，默认值为 10249
--hostname-override string	设置本 Node 在集群中的主机名，不设置将使用本机 hostname
--iptables-masquerade-bit int32	标记数据包将进行 SNAT 的 fwmark 位设置，有效范围为[0, 31]，默认值为 14
--iptables-min-sync-period duration	刷新 Iptables 规则的最小时间间隔，例如 5s、1m、2h22m
--iptables-sync-period duration	刷新 Iptables 规则的时间间隔，例如 5s、1m、2h22m，默认值为 30s，必须大于 0
--kube-api-burst int32	每秒发送到 API Server 的请求的数量，默认值为 10
--kube-api-content-type string	发送到 API Server 的请求内容类型，默认值为 application/vnd.kubernetes.protobuf
--kube-api-qps float32	与 API Server 通信的 QPS 值，默认值为 5
--kubeconfig string	kubeconfig 配置文件路径，在配置文件中包括 Master 地址信息及必要的认证信息
--masquerade-all	设置为 true 时表示使用纯 Iptables 代理，所有网络包都将做 SNAT 转换
--master string	API Server 的地址
--oom-score-adj int32	kube-proxy 进程的 oom_score_adj 参数值，有效范围为[-1000,1000]，默认值为 -999
--proxy-mode ProxyMode	代理模式，可选项为 iptables 或 userspace，默认值为 iptables，转发速度更快。当操作系统 kernel 版本或 Iptables 版本不够新时，将自动降级为 userspace 模式
--proxy-port-range port-range	进行 Service 代理的本地端口号范围，格式为 begin-end，含两端，未指定则采用随机选择的系统可用的端口号
--udp-timeout duration	保持空闲 UDP 连接的时间，例如 250ms、2s，必须大于 0，仅当 proxy-mode=userspace 时生效，默认值为 250ms

## 2.2 kubectl 命令行工具用法详解

kubectl 作为客户端 CLI 工具，可以让用户通过命令行的方式对 Kubernetes 集群进行操作。本节对 kubectl 的子命令和用法进行详细说明。

### 2.2.1 kubectl 用法概述

kubectl 命令行的语法如下：

```
$ kubectl [command] [TYPE] [NAME] [flags]
```

其中，command、TYPE、NAME、flags 的含义如下。

(1) **command**：子命令，用于操作 Kubernetes 集群资源对象的命令，例如 create、delete、describe、get、apply 等。

(2) **TYPE**：资源对象的类型，区分大小写，能以单数形式、复数形式或者简写形式表示。例如以下 3 种 TYPE 是等价的。

```
$ kubectl get pod pod1
$ kubectl get pods pod1
$ kubectl get po pod1
```

(3) **NAME**：资源对象的名称，区分大小写。如果不指定名称，则系统将返回属于 TYPE 的全部对象的列表，例如 \$ kubectl get pods 将返回所有 Pod 的列表。

(4) **flags**：kubectl 子命令的可选参数，例如使用 “-s” 指定 apiserver 的 URL 地址而不用默认值。

kubectl 可操作的资源对象类型如表 2.9 所示。

表 2.9 kubectl 可操作的资源对象类型

资源对象的名称	缩 写
clusters	
componentstatuses	cs
configmaps	cm
daemonsets	ds
deployments	deploy
endpoints	ep
events	ev
horizontalpodautoscalers	hpa



续表

资源对象的名称	缩 写
ingresses	ing
Jobs	
limitranges	limits
nodes	no
namespaces	ns
networkpolicies	
nodes	no
statefulsets	
persistentvolumeclaims	pvc
persistentvolumes	pv
Pods	po
podsecuritypolicies	psp
podtemplates	
replicasets	rs
replicationcontrollers	rc
resourcequotas	quota
cronjob	
secrets	
serviceaccounts	
services	svc
storageclasses	sc
thirdpartyresources	

在一个命令行中也可以同时对多个资源对象进行操作,以多个 TYPE 和 NAME 的组合表示,示例如下。

◎ 获取多个 Pod 的信息:

```
$ kubectl get pods pod1 pod2
```

◎ 获取多种对象的信息:

```
$ kubectl get pod/pod1 rc/rc1
```

◎ 同时应用多个 yaml 文件,以多个 -f file 参数表示:

```
$ kubectl get pod -f pod1.yaml -f pod2.yaml
$ kubectl create -f pod1.yaml -f rc1.yaml -f service1.yaml
```

## 2.2.2 kubectl 子命令详解

kubectl 的子命令非常丰富，涵盖了对 Kubernetes 集群的主要操作，包括资源对象的创建、删除、查看、修改、配置、运行等。详细的子命令如表 2.10 所示。

表 2.10 kubectl 子命令详解

子 命 令	语 法	说 明
annotate	kubectl annotate (-f FILENAME   TYPE NAME   TYPE/NAME) KEY_1=VAL_1 ... KEY_N=VAL_N [--overwrite] [--all] [--resource-version=version] [flags]	添加或更新资源对象的 annotation 信息
api-versions	kubectl api-versions [flags]	列出当前系统支持的 API 版本列表，格式为“group/version”
apply	kubectl apply -f FILENAME [flags]	从配置文件或 stdin 中对资源对象进行配置更新
attach	kubectl attach POD -c CONTAINER [flags]	附着到一个正在运行的容器上
autoscale	kubectl autoscale (-f FILENAME   TYPE NAME   TYPE/NAME) [--min=MINPODS] --max=MAXPODS [--cpu-percent=CPU] [flags]	对 Deployment、ReplicaSet 或 ReplicationController 进行水平自动扩容和缩容的设置
cluster-info	kubectl cluster-info [flags]	显示集群 Master 和内置服务的信息
completion	kubectl completion SHELL [flags]	输出 shell 命令的执行结果码（bash 或 zsh）
config	kubectl config SUBCOMMAND [flags]	修改 kubeconfig 文件
convert	kubectl convert -f FILENAME [flags]	转换配置文件为不同的 API 版本
cordon	kubectl cordon NODE [flags]	将 Node 标记为 unschedulable，即“隔离”出集群调度范围
create	kubectl create -f FILENAME [flags]	从配置文件或 stdin 中创建资源对象
delete	kubectl delete (-f FILENAME   TYPE [NAME   /NAME   -l label   --all]) [flags]	根据配置文件、stdin、资源名称或 label selector 删除资源对象
describe	kubectl describe (-f FILENAME   TYPE [NAME_PREFIX   /NAME   -l label]) [flags]	描述一个或多个资源对象的详细信息
drain	kubectl drain NODE [flags]	首先将 Node 设置为 unschedulable，然后删除该 Node 上运行的所有 Pod，但不会删除不由 apiserver 管理的 Pod
edit	kubectl edit (-f FILENAME   TYPE NAME   TYPE/NAME) [flags]	编辑资源对象的属性，在线更新
exec	kubectl exec POD [-c CONTAINER] [-i] [-t] [flags] [-- COMMAND [args...]]	执行一个容器中的命令
explain	kubectl explain [--include-extended-apis=true] [--recursive=false] [flags]	对资源对象属性的详细说明

续表

子 命 令	语 法	说 明
expose	kubectl expose (-f FILENAME   TYPE NAME   TYPE/NAME) [--port=port] [--protocol=TCP UDP] [--target-port=number-or-name] [--name=name] [--external-ip=external-ip-of-service] [--type=type] [flags]	将已经存在的一个 RC、Service、Deployment 或 Pod 暴露为一个新的 Service
get	kubectl get (-f FILENAME   TYPE [NAME   /NAME   -l label]) [--watch] [--sort-by=FIELD] [--o   --output]=OUTPUT_FORMAT [flags]	显示一个或多个资源对象的概要信息
label	kubectl label (-f FILENAME   TYPE NAME   TYPE/NAME) KEY_1=VAL_1 ... KEY_N=VAL_N [--overwrite] [--all] [--resource-version=version] [flags]	设置或更新资源对象的 labels
logs	kubectl logs POD [-c CONTAINER] [--follow] [flags]	在屏幕上打印一个容器的日志
Patch	kubectl patch (-f FILENAME   TYPE NAME   TYPE/NAME) --patch PATCH [flags] kubectl patch (-f FILENAME   TYPE NAME   TYPE/NAME) --patch PATCH [flags]	以 merge 形式对资源对象的部分字段的值进行修改
port-forward	kubectl port-forward POD [LOCAL_PORT:]REMOTE_PORT [...[LOCAL_PORT_N:]REMOTE_PORT_N] [flags]	将本机的某个端口号映射到 Pod 的端口号，通常用于测试工作
proxy	kubectl proxy [--port=PORT] [--www=static-dir] [--www-prefix=prefix] [--api-prefix=prefix] [flags]	将本机某个端口号映射到 apiserver
replace	kubectl replace -f FILENAME [flags]	从配置文件或 stdin 替换资源对象
rolling-update	kubectl rolling-update OLD_CONTROLLER_NAME ([NEW_CONTROLLER_NAME] --image=NEW_CONTAINER_IMAGE   -f NEW_CONTROLLER_SPEC) [flags]	对 RC 进行滚动升级
rollout	kubectl rollout SUBCOMMAND [flags]	对 Deployment 进行管理，可用操作包括: history、pause、resume、undo、status
run	kubectl run NAME --image=image [--env="key=value"] [--port=port] [--replicas=replicas] [--dry-run=bool] [--overrides=inline-json] [flags]	基于一个镜像在 Kubernetes 集群上启动一个 Deployment
scale	kubectl scale (-f FILENAME   TYPE NAME   TYPE/NAME) --replicas=COUNT [--resource-version=version] [--current-replicas=count] [flags]	扩容、缩容一个 Deployment、ReplicaSet、RC 或 Job 中 Pod 的数量
set	kubectl set SUBCOMMAND [flags]	设置资源对象的某个特定信息，目前仅支持修改容器的镜像
taint	kubectl taint NODE NAME KEY_1=VAL_1:TAINT_EFFECT_1 ... KEY_N=VAL_N:TAINT_EFFECT_N [flags]	设置 Node 的 taint 信息，用于将特定的 Pod 调度到特定的 Node 的操作，为 Alpha 版本的功能
uncordon	kubectl uncordon NODE [flags]	将 Node 设置为 schedulable
version	kubectl version [--client] [flags]	打印系统的版本信息

### 2.2.3 kubectl 参数列表

kubectl 命令行的公共启动参数如表 2.11 所示。

表 2.11 kubectl 命令行的公共启动参数

参数名和取值示例	说 明
--alsologtostderr=false	设置为 true 表示将日志输出到文件的同时输出到 stderr
--as=""	设置本次操作的用户名
--certificate-authority="	用于 CA 授权的 cert 文件路径
--client-certificate="	用于 TLS 的客户端证书文件路径
--client-key="	用于 TLS 的客户端 key 文件路径
--cluster="	设置要使用的 kubeconfig 中的 cluster 名
--context="	设置要使用的 kubeconfig 中的 context 名
--insecure-skip-tls-verify=false	设置为 true 表示跳过 TLS 安全验证模式，将使得 HTTPS 连接不安全
--kubeconfig="	kubeconfig 配置文件路径，在配置文件中包括 Master 地址信息及必要的认证信息
--log-backtrace-at=:0	记录日志每到“file:行号”时打印一次 stack trace
--log-dir="	日志文件路径
--log-flush-frequency=5s	设置 flush 日志文件的时间间隔
--logtostderr=true	设置为 true 表示将日志输出到 stderr，不输出到日志文件
--match-server-version=false	设置为 true 表示客户端版本号需要与服务端一致
--namespace="	设置本次操作所在的 namespace
--password="	设置 apiserver 的 basic authentication 的密码
-s, --server="	设置 apiserver 的 URL 地址，默认值为 localhost:8080
--stderrthreshold=2	在该 threshold 级别之上的日志将输出到 stderr
--token="	设置访问 apiserver 的安全 token
--user="	指定 kubeconfig 用户名
--username="	设置 apiserver 的 basic authentication 的用户名
--v=0	glog 日志级别
--vmodule=	glog 基于模块的详细日志级别

每个子命令（如 create、delete、get 等）还有特定的 flags 参数，可以通过\$ kubectl [command] --help 命令进行查看。

### 2.2.4 kubectl 输出格式

kubectl 命令可以用多种格式对结果进行显示，输出的格式通过-o 参数指定：

```
$ kubectl [command] [TYPE] [NAME] -o=<output_format>
```

根据不同子命令的输出结果，可选的输出格式如表 2.12 所示。

表 2.12 kubectl 命令的输出格式列表

输出格式	说 明
-o=custom-columns=<spec>	根据自定义列名进行输出，以逗号分隔
-o=custom-columns-file=<filename>	从文件中获取自定义列名进行输出
-o=json	以 JSON 格式显示结果
-o=jsonpath=<template>	输出 jsonpath 表达式定义的字段信息
-o=jsonpath-file=<filename>	输出 jsonpath 表达式定义的字段信息，来源于文件
-o=name	仅输出资源对象的名称
-o=wide	输出额外信息。对于 Pod，将输出 Pod 所在的 Node 名
-o=yaml	以 yaml 格式显示结果

常用的输出格式示例如下。

(1) 显示 Pod 的更多信息：

```
$ kubectl get pod <pod-name> -o wide
```

(2) 以 yaml 格式显示 Pod 的详细信息：

```
$ kubectl get pod <pod-name> -o yaml
```

(3) 以自定义列名显示 Pod 的信息：

```
$ kubectl get pod <pod-name>
-o=custom-columns=NAME:.metadata.name,RSRC:.metadata.resourceVersion
```

(4) 基于文件的自定义列名输出：

```
$ kubectl get pods <pod-name> -o=custom-columns-file=template.txt
```

template.txt 文件的内容为：

```
NAME          RSRC
metadata.name  metadata.resourceVersion
```

输出结果为：

```
NAME      RSRC
pod-name  52305
```

另外，还可以将输出结果按某个字段排序，通过--sort-by 参数以 jsonpath 表达式进行指定：

```
$ kubectl [command] [TYPE] [NAME] --sort-by=<jsonpath_exp>
```

例如，按照名字进行排序：

```
$ kubectl get pods --sort-by=.metadata.name
```

## 2.2.5 kubectl 操作示例

---

本节将一些常用的 kubectl 操作作为示例进行说明。

### 1. 创建资源对象

根据 yaml 配置文件一次性创建 service 和 rc：

```
$ kubectl create -f my-service.yaml -f my-rc.yaml
```

根据<directory>目录下所有.yaml、.yml、.json 文件的定义进行创建操作：

```
$ kubectl create -f <directory>
```

### 2. 查看资源对象

查看所有 Pod 列表：

```
$ kubectl get pods
```

查看 rc 和 service 列表：

```
$ kubectl get rc,service
```

### 3. 描述资源对象

显示 Node 的详细信息：

```
$ kubectl describe nodes <node-name>
```

显示 Pod 的详细信息：

```
$ kubectl describe pods/<pod-name>
```

显示由 RC 管理的 Pod 的信息：

```
$ kubectl describe pods <rc-name>
```

### 4. 删除资源对象

基于 pod.yaml 定义的名称删除 Pod：

```
$ kubectl delete -f pod.yaml
```

删除所有包含某个 label 的 Pod 和 service：

```
$ kubectl delete pods,services -l name=<label-name>
```

删除所有 Pod：

```
$ kubectl delete pods --all
```

## 5. 执行容器的命令

执行 Pod 的 `date` 命令，默认使用 Pod 中的第 1 个容器执行：

```
$ kubectl exec <pod-name> date
```

指定 Pod 中某个容器执行 `date` 命令：

```
$ kubectl exec <pod-name> -c <container-name> date
```

通过 `bash` 获得 Pod 中某个容器的 TTY，相当于登录容器：

```
$ kubectl exec -ti <pod-name> -c <container-name> /bin/bash
```

## 6. 查看容器的日志

查看容器输出到 `stdout` 的日志：

```
$ kubectl logs <pod-name>
```

跟踪查看容器的日志，相当于 `tail -f` 命令的结果：

```
$ kubectl logs -f <pod-name> -c <container-name>
```

## 2.3 深入掌握 Pod

接下来，让我们深入探索 Pod 的应用、配置、调度管理及服务的应用，开始 Kubernetes 容器编排之旅。

本节将对 Kubernetes 如何发布和管理应用进行详细说明和示例，主要包括 Pod 和容器的使用、Pod 的控制和调度管理、应用配置管理等内容。

### 2.3.1 Pod 定义详解

yaml 格式的 Pod 定义文件的完整内容如下：

```
apiVersion: v1
kind: Pod
metadata:
  name: string
  namespace: string
  labels:
    - name: string
  annotations:
    - name: string
```

```
spec:
  containers:
  - name: string
    image: string
    imagePullPolicy: [Always | Never | IfNotPresent]
    command: [string]
    args: [string]
    workingDir: string
    volumeMounts:
    - name: string
      mountPath: string
      readOnly: boolean
    ports:
    - name: string
      containerPort: int
      hostPort: int
      protocol: string
    env:
    - name: string
      value: string
    resources:
      limits:
        cpu: string
        memory: string
      requests:
        cpu: string
        memory: string
    livenessProbe:
      exec:
        command: [string]
      httpGet:
        path: string
        port: number
        host: string
        scheme: string
        httpHeaders:
        - name: string
          value: string
      tcpSocket:
        port: number
    initialDelaySeconds: 0
    timeoutSeconds: 0
    periodSeconds: 0
    successThreshold: 0
```



```

    failureThreshold: 0
  securityContext:
    privileged: false
  restartPolicy: [Always | Never | OnFailure]
  nodeSelector: object
  imagePullSecrets:
  - name: string
  hostNetwork: false
  volumes:
  - name: string
    emptyDir: {}
    hostPath:
      path: string
    secret:
      secretName: string
      items:
      - key: string
        path: string
  configMap:
    name: string
    items:
    - key: string
      path: string

```

对各属性的详细说明如表 2.13 所示。

表 2.13 对 Pod 定义文件模板中各属性的详细说明

属 性 名 称	取 值 类 型	是 否 必 选	取 值 说 明
version	String	Required	版本号, 例如 v1
kind	String	Required	Pod
metadata	Object	Required	元数据
metadata.name	String	Required	Pod 的名称, 命名规范需符合 RFC 1035 规范
metadata.namespace	String	Required	Pod 所属的命名空间, 默认值为 default
metadata.labels[]	List		自定义标签列表
metadata.annotation[]	List		自定义注解列表
Spec	Object	Required	Pod 中容器的详细定义
spec.containers[]	List	Required	Pod 中的容器列表
spec.containers[].name	String	Required	容器的名称, 需符合 RFC 1035 规范
spec.containers[].image	String	Required	容器的镜像名称

续表

属 性 名 称	取 值 类 型	是 否 必 选	取 值 说 明
spec.containers[].imagePullPolicy	String		获取镜像的策略，可选值包括：Always、Never、IfNotPresent，默认值为 Always。 Always：表示每次都尝试重新下载镜像。 IfNotPresent：表示如果本地有该镜像，则使用本地的镜像，本地不存在时下载镜像。 Never：表示仅使用本地镜像
spec.containers[].command[]	List		容器的启动命令列表，如果不指定，则使用镜像打包时使用的启动命令
spec.containers[].args[]	List		容器的启动命令参数列表
spec.containers[].workingDir	String		容器的工作目录
spec.containers[].volumeMounts[]	List		挂载到容器内部的存储卷配置
spec.containers[].volumeMounts[].name	String		引用 Pod 定义的共享存储卷的名称，需使用 volumes[]部分定义的共享存储卷名称
spec.containers[].volumeMounts[].mountPath	String		存储卷在容器内 Mount 的绝对路径，应少于 512 个字符
spec.containers[].volumeMounts[].readOnly	Boolean		是否为只读模式，默认值为读写模式
spec.containers[].ports[]	List		容器需要暴露的端口号列表
spec.containers[].ports[].name	String		端口的名称
spec.containers[].ports[].containerPort	Int		容器需要监听的端口号
spec.containers[].ports[].hostPort	Int		容器所在主机需要监听的端口号，默认与 containerPort 相同。设置 hostPort 时，同一台宿主主机将无法启动该容器的第 2 份副本
spec.containers[].ports[].protocol	String		端口协议，支持 TCP 和 UDP，默认值为 TCP
spec.containers[].env[]	List		容器运行前需设置的环境变量列表
spec.containers[].env[].name	String		环境变量的名称
spec.containers[].env[].value	String		环境变量的值
spec.containers[].resources	Object		资源限制和资源请求的设置，详见第 5 章的说明
spec.containers[].resources.limits	Object		资源限制的设置
spec.containers[].resources.limits.cpu	String		CPU 限制，单位为 core 数，将用于 docker run --cpu-shares 参数
spec.containers[].resources.limits.memory	String		内存限制，单位可以为 MiB/GiB 等，将用于 docker run --memory 参数
spec.containers[].resources.requests	Object		资源限制的设置

续表

属 性 名 称	取 值 类 型	是 否 必 选	取 值 说 明
spec.containers[].resources.requests.cpu	String		CPU 请求, 单位为 core 数, 容器启动的初始可用数量
spec.containers[].resources.requests.memory	String		内存请求, 单位可以为 MiB、GiB 等, 容器启动的初始可用数量
spec.volumes[]	List		在该 Pod 上定义的共享存储卷列表
spec.volumes[].name	String		共享存储卷的名称, 在一个 Pod 中每个存储卷定义一个名称, 应符合 RFC 1035 规范。容器定义部分的 containers[].volumeMounts[].name 将引用该共享存储卷的名称。  Volume 的类型包括: emptyDir、hostPath、gcePersistentDisk、awsElasticBlockStore、gitRepo、secret、nfs、iscsi、glusterfs、persistentVolumeClaim、rbd、flexVolume、cinder、cephfs、flocker、downwardAPI、fc、azureFile、configMap、vsphereVolume, 可以定义多个 volume, 每个 volume 的 name 保持唯一。本节讲解 emptyDir、hostPath、secret、configMap 这 4 种 volume, 其他类型 volume 的设置方式详见第 1 章的说明
spec.volumes[].emptyDir	Object		类型为 emptyDir 的存储卷, 表示与 Pod 同生命周期的一个临时目录, 其值为一个空对象: emptyDir: {}
spec.volumes[].hostPath	Object		类型为 hostPath 的存储卷, 表示挂载 Pod 所在宿主机的目录, 通过 volumes[].hostPath.path 指定
spec.volumes[].hostPath.path	String		Pod 所在主机的目录, 将被用于容器中 mount 的目录
spec.volumes[].secret	Object		类型为 secret 的存储卷, 表示挂载集群预定义的 secret 对象到容器内部
spec.volumes[].configMap	Object		类型为 configMap 的存储卷, 表示挂载集群预定义的 configMap 对象到容器内部
spec.volumes[].livenessProbe	Object		对 Pod 内各容器健康检查的设置, 当探测无响应几次之后, 系统将自动重启该容器。可以设置的方法包括: exec、httpGet 和 tcpSocket。对一个容器仅需设置一种健康检查方法
spec.volumes[].livenessProbe.exec	Object		对 Pod 内各容器健康检查的设置, exec 方式
spec.volumes[].livenessProbe.exec.command[]	String		exec 方式需要指定的命令或者脚本
spec.volumes[].livenessProbe.httpGet	Object		对 Pod 内各容器健康检查的设置, HTTPGet 方式。需指定 path、port

续表

属 性 名 称	取 值 类 型	是 否 必 选	取 值 说 明
spec.volumes[].livenessProbe.tcpSocket	Object		对 Pod 内各容器健康检查的设置，tcpSocket 方式
spec.volumes[].livenessProbe.initialDelaySeconds	Number		容器启动完成后进行首次探测的时间，单位为 s
spec.volumes[].livenessProbe.timeoutSeconds	Number		对容器健康检查的探测等待响应的超时时间设置，单位为 s，默认值为 1s。超过该超时时间设置，将认为该容器不健康，将重启该容器
spec.volumes[].livenessProbe.periodSeconds	Number		对容器健康检查的定期探测时间设置，单位为 s，默认为 10s 探测一次
spec.restartPolicy	String		Pod 的重启策略，可选值为 Always、OnFailure，默认值为 Always。 Always: Pod 一旦终止运行，则无论容器是如何终止的，kubelet 都将重启它。 OnFailure: 只有 Pod 以非零退出码终止时，kubelet 才会重启该容器。如果容器正常结束（退出码为 0），则 kubelet 将不会重启它。 Never: Pod 终止后，kubelet 将退出码报告给 Master，不会再重启该 Pod
spec.nodeSelector	Object		设置 NodeSelector 表示将该 Pod 调度到包含这些 label 的 Node 上，以 key:value 格式指定
spec.imagePullSecrets	Object		Pull 镜像时使用的 secret 名称，以 name:secretkey 格式指定
spec.hostNetwork	Boolean		是否使用主机网络模式，默认值为 false。如果设置为 true，则表示容器使用宿主机网络，不再使用 Docker 网桥，该 Pod 将无法在同一台宿主机上启动第 2 个副本

2.3.2 Pod 的基本用法

在对 Pod 的用法进行说明之前，有必要先对 Docker 容器中应用的运行要求进行说明。

在使用 Docker 时，可以使用 docker run 命令创建并启动一个容器。而在 Kubernetes 系统中对长时间运行容器的要求是：其主程序需要一直在前台执行。如果我们创建的 Docker 镜像的启动命令是后台执行程序，例如 Linux 脚本：

```
nohup ./start.sh &
```

则在 kubelet 创建包含这个容器的 Pod 之后运行完该命令，即认为 Pod 执行结束，将立刻销毁该 Pod。如果为该 Pod 定义了 ReplicationController，则系统将会监控到该 Pod 已经终止，之后根

据 RC 定义中 Pod 的 replicas 副本数量生成一个新的 Pod。而一旦创建出新的 Pod，就将在执行完启动命令后，陷入无限循环的过程中。这就是 Kubernetes 需要我们自己创建的 Docker 镜像以一个前台命令作为启动命令的原因。

对于无法改造为前台执行的应用，也可以使用开源工具 Supervisor 辅助进行前台运行的功能。Supervisor 提供了一种可以同时启动多个后台应用，并保持 Supervisor 自身在前台执行的机制，可以满足 Kubernetes 对容器的启动要求。关于 Supervisor 的安装和使用，请参考官网 <http://supervisord.org> 的文档说明。

接下来对 Pod 对容器的封装和应用进行说明，Pod 的基本用法为：Pod 可以由 1 个或多个容器组合而成。

在上一节 Guestbook 的例子中，名为 frontend 的 Pod 只由一个容器组成：

```
apiVersion: v1
kind: Pod
metadata:
  name: frontend
  labels:
    name: frontend
spec:
  containers:
  - name: frontend
    image: kubeguide/guestbook-php-frontend
    env:
    - name: GET_HOSTS_FROM
      value: env
    ports:
    - containerPort: 80
```

这个 frontend Pod 在成功启动之后，将启动 1 个 Docker 容器。

另一种场景是，当 frontend 和 redis 两个容器应用为紧耦合的关系，应该组合成一个整体对外提供服务时，应将这两个容器打包为一个 Pod，如图 2.2 所示。

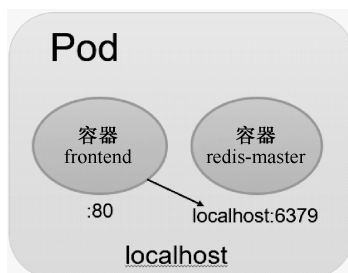


图 2.2 包含两个容器的 Pod

配置文件 frontend-localredis-pod.yaml 如下：

```
apiVersion: v1
kind: Pod
metadata:
  name: redis-php
  labels:
    name: redis-php
spec:
  containers:
    - name: frontend
      image: kubeguide/guestbook-php-frontend:localredis
      ports:
        - containerPort: 80
    - name: redis
      image: kubeguide/redis-master
      ports:
        - containerPort: 6379
```

属于一个 Pod 的多个容器应用之间相互访问时仅需要通过 localhost 就可以通信，使得这一组容器被“绑定”在了一个环境中。

在 Docker 容器 kubeguide/guestbook-php-frontend:localredis 的 PHP 网页中，直接通过 URL 地址“localhost:6379”对同属于一个 Pod 内的 redis-master 进行访问。guestbook.php 的内容如下：

```
<?
set_include_path('..usr/local/lib/php');
error_reporting(E_ALL);
ini_set('display_errors', 1);
require 'Predis/Autoloader.php';
Predis\Autoloader::register();

if (isset($_GET['cmd']) === true) {
    $host = 'localhost';
    if (getenv('REDIS_HOST') && strlen(getenv('REDIS_HOST')) > 0 ) {
        $host = getenv('REDIS_HOST');
    }
    header('Content-Type: application/json');
    if ($_GET['cmd'] == 'set') {
        $client = new Predis\Client([
            'scheme' => 'tcp',
            'host'    => $host,
            'port'    => 6379,
        ]);

        $client->set($_GET['key'], $_GET['value']);
        print("{\"message": "Updated"}");
    } else {
```

```

$host = 'localhost';
if (getenv('REDIS_HOST') && strlen(getenv('REDIS_HOST')) > 0 ) {
    $host = getenv('REDIS_HOST');
}
$client = new Predis\Client([
    'scheme' => 'tcp',
    'host'    => $host,
    'port'    => 6379,
]);

$value = $client->get($_GET['key']);
print('{"data": "' . $value . '"}');
}
} else {
    phpinfo();
} ?>

```

运行 `kubectl create` 命令创建该 Pod:

```

$ kubectl create -f frontend-localredis-pod.yaml
pod "redis-php" created

```

查看已创建的 Pod:

```

# kubectl get pods
NAME          READY   STATUS    RESTARTS   AGE
redis-php     2/2     Running   0           10m

```

可以看到 **READY** 信息为 2/2，表示 Pod 中的两个容器都成功运行了。

查看这个 Pod 的详细信息，可以看到两个容器的定义及创建的过程（Event 事件信息）:

```

# kubectl describe pod redis-php
Name:          redis-php
Namespace:     default
Node:          k8s/192.168.18.3
Start Time:    Thu, 28 Jul 2016 12:28:21 +0800
Labels:        name=redis-php
Status:        Running
IP:            172.17.1.4
Controllers:   <none>
Containers:
  frontend:
    Container ID:
docker://ccc8616f8df1fb19abbd0ab189a36e6f6628b78ba7b97b1077d86e7fc224ee08
    Image:          kubeguide/guestbook-php-frontend:localredis
    Image ID:
docker://sha256:d014f67384a11186e135b95a7ed0d794674f7ce258f0dce47267c3052a0d0fa9
    Port:          80/TCP
    State:          Running

```

```

    Started:                Thu, 28 Jul 2016 12:28:22 +0800
    Ready:                  True
    Restart Count:          0
    Environment Variables:  <none>
  redis:
    Container ID:
docker://c0b19362097cda6dd5b8ed7d8eaaaf43aeeb969ee023ef255604bde089808075
    Image:                  kubeguide/redis-master
    Image ID:
docker://sha256:405a0b586f7ebef545ec65be0e914311159d1baedccd3a93e9d3e3b249ec5cbd
    Port:                   6379/TCP
    State:                  Running
      Started:              Thu, 28 Jul 2016 12:28:23 +0800
    Ready:                  True
    Restart Count:          0
    Environment Variables:  <none>
Conditions:
  Type      Status
  Initialized True
  Ready      True
  PodSchedul True
Volumes:
  default-token-97j21:
    Type:          Secret (a volume populated by a Secret)
    SecretName:    default-token-97j21
QoS Tier:        BestEffort
Events:
  FirstSeen    LastSeen    Count    From          SubobjectPath  Type    Reason    Message
  -----
  18m          18m         1        {default-scheduler }    Normal
Scheduled      Successfully assigned redis-php to k8s-node-1
  18m          18m         1        {kubelet k8s-node-1}
spec.containers{frontend}    Normal    Pulled    Container image
"kubeguide/guestbook-php-frontend:localredis" already present on machine
  18m          18m         1        {kubelet k8s-node-1}
spec.containers{frontend}    Normal    Created   Created container
with docker id ccc8616f8df1
  18m          18m         1        {kubelet k8s-node-1}
spec.containers{frontend}    Normal    Started   Started container
with docker id ccc8616f8df1
  18m          18m         1        {kubelet k8s-node-1}
spec.containers{redis}       Normal    Pulled    Container image
"kubeguide/redis-master" already present on machine
  18m          18m         1        {kubelet k8s-node-1}
spec.containers{redis}       Normal    Created   Created container
with docker id c0b19362097c
  18m          18m         1        {kubelet k8s-node-1}
```



```
spec.containers{redis}           Normal          Started          Started container
with docker id c0b19362097c
```

### 2.3.3 静态 Pod

静态 Pod 是由 kubelet 进行管理的仅存在于特定 Node 上的 Pod。它们不能通过 API Server 进行管理，无法与 ReplicationController、Deployment 或者 DaemonSet 进行关联，并且 kubelet 也无法对它们进行健康检查。静态 Pod 总是由 kubelet 进行创建的，并且总是在 kubelet 所在的 Node 上运行的。

创建静态 Pod 有两种方式：配置文件方式和 HTTP 方式。

#### 1) 配置文件方式

首先，需要设置 kubelet 的启动参数 “--config”，指定 kubelet 需要监控的配置文件所在的目录，kubelet 会定期扫描该目录，并根据该目录中的 .yaml 或 .json 文件进行创建操作。

假设配置目录为 /etc/kubelet.d/，配置启动参数：--config=/etc/kubelet.d/，然后重启 kubelet 服务。

在目录 /etc/kubelet.d 中放入 static-web.yaml 文件，内容如下：

```
apiVersion: v1
kind: Pod
metadata:
  name: static-web
  labels:
    name: static-web
spec:
  containers:
  - name: static-web
    image: nginx
    ports:
    - name: web
      containerPort: 80
```

等待一会儿，查看本机中已经启动的容器：

```
# docker ps
CONTAINER ID   IMAGE     COMMAND                  CREATED        STATUS        PORTS          NAMES
2292ea231ab1   nginx    "nginx -g 'daemon off'" 1 minute ago   1m
k8s_static-web.68ee0075_static-web-k8s-node-1_default_78c7efddebf191c949cbb7aa22a927c8_401b96d0
```

可以看到一个 Nginx 容器已经被 kubelet 成功创建了出来。

到 Master 节点查看 Pod 列表，可以看到这个 static pod：

```
# kubectl get pods
```

NAME	READY	STATUS	RESTARTS	AGE
static-web-node1	1/1	Running	0	5m

由于静态 Pod 无法通过 API Server 直接管理，所以在 Master 节点尝试删除这个 Pod，将会使其变成 Pending 状态，且不会被删除。

```
# kubectl delete pod static-web-node1
pod "static-web-node1" deleted

# kubectl get pods
NAME          READY   STATUS    RESTARTS   AGE
static-web-node1 0/1     Pending   0          1s
```

删除该 Pod 的操作只能是到其所在 Node 上，将其定义文件 static-web.yaml 从/etc/kubelet.d 目录下删除。

```
# rm /etc/kubelet.d/static-web.yaml
# docker ps
// 无容器正在运行。
```

2) HTTP 方式

通过设置 kubelet 的启动参数 “--manifest-url”，kubelet 将会定期从该 URL 地址下载 Pod 的定义文件，并以.yaml 或.json 文件的格式进行解析，然后创建 Pod。其实现方式与配置文件方式是一致的。

2.3.4 Pod 容器共享 Volume

在同一个 Pod 中的多个容器能够共享 Pod 级别的存储卷 Volume。Volume 可以被定义为各种类型，多个容器各自进行挂载操作，将一个 Volume 挂载为容器内部需要的目录，如图 2.3 所示。

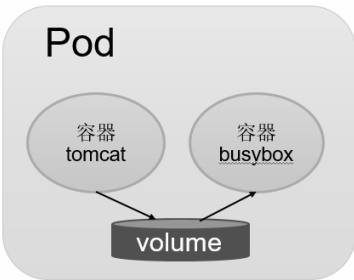


图 2.3 Pod 中多个容器共享 volume

在下面的例子中，Pod 内包含两个容器：tomcat 和 busybox，在 Pod 级别设置 Volume “app-logs”，用于 tomcat 向其中写日志文件，busybox 读日志文件。

配置文件 pod-volume-applogs.yaml 的内容如下:

```
apiVersion: v1
kind: Pod
metadata:
  name: volume-pod
spec:
  containers:
  - name: tomcat
    image: tomcat
    ports:
    - containerPort: 8080
    volumeMounts:
    - name: app-logs
      mountPath: /usr/local/tomcat/logs
  - name: busybox
    image: busybox
    command: ["sh", "-c", "tail -f /logs/catalina*.log"]
    volumeMounts:
    - name: app-logs
      mountPath: /logs
  volumes:
  - name: app-logs
    emptyDir: {}
```

这里设置的 Volume 名为 app-logs，类型为 emptyDir（也可以设置为其他类型，详见第1章对 Volume 概念的说明），挂载到 tomcat 容器内的 /usr/local/tomcat/logs 目录，同时挂载到 logreader 容器内的 /logs 目录。tomcat 容器在启动后会向 /usr/local/tomcat/logs 目录中写文件，logreader 容器就可以读取其中的文件了。

logreader 容器的启动命令为 tail -f /logs/catalina\*.log，我们可以通过 kubectl logs 命令查看 logreader 容器的输出内容：

```
# kubectl logs volume-pod -c busybox
.....
29-Jul-2016 12:55:59.626 INFO [localhost-startStop-1]
org.apache.catalina.startup.HostConfig.deployDirectory Deploying web application
directory /usr/local/tomcat/webapps/manager
29-Jul-2016 12:55:59.722 INFO [localhost-startStop-1]
org.apache.catalina.startup.HostConfig.deployDirectory Deployment of web
application directory /usr/local/tomcat/webapps/manager has finished in 96 ms
29-Jul-2016 12:55:59.740 INFO [main] org.apache.coyote.AbstractProtocol.start
Starting ProtocolHandler ["http-apr-8080"]
29-Jul-2016 12:55:59.794 INFO [main] org.apache.coyote.AbstractProtocol.start
Starting ProtocolHandler ["ajp-apr-8009"]
29-Jul-2016 12:56:00.604 INFO [main] org.apache.catalina.startup.Catalina.start
Server startup in 4052 ms
```

这个文件即为 tomcat 生成的日志文件/usr/local/tomcat/logs/catalina.<date>.log 的内容。登录 tomcat 容器进行查看：

```
# kubectl exec -ti volume-pod -c tomcat -- ls /usr/local/tomcat/logs
catalina.2016-07-29.log      localhost_access_log.2016-07-29.txt
host-manager.2016-07-29.log manager.2016-07-29.log

# kubectl exec -ti volume-pod -c tomcat -- tail
/usr/local/tomcat/logs/catalina.2016-07-29.log
.....
29-Jul-2016 12:55:59.722 INFO [localhost-startStop-1]
org.apache.catalina.startup.HostConfig.deployDirectory Deployment of web
application directory /usr/local/tomcat/webapps/manager has finished in 96 ms
29-Jul-2016 12:55:59.740 INFO [main] org.apache.coyote.AbstractProtocol.start
Starting ProtocolHandler ["http-apr-8080"]
29-Jul-2016 12:55:59.794 INFO [main] org.apache.coyote.AbstractProtocol.start
Starting ProtocolHandler ["ajp-apr-8009"]
29-Jul-2016 12:56:00.604 INFO [main] org.apache.catalina.startup.Catalina.start
Server startup in 4052 ms
```

### 2.3.5 Pod 的配置管理

应用部署的一个最佳实践是将应用所需的配置信息与程序进行分离，这样可以使得应用程序被更好地复用，通过不同的配置也能实现更灵活的功能。将应用打包为容器镜像后，可以通过环境变量或者外挂文件的方式在创建容器时进行配置注入，但在大规模容器集群的环境中，对多个容器进行不同的配置将变得非常复杂。从 Kubernetes v1.2 开始提供了一种统一的应用配置管理方案——ConfigMap。本节对 ConfigMap 的概念和用法进行详细描述。

#### 1. ConfigMap 概述

ConfigMap 供容器使用的典型用法如下。

- (1) 生成为容器内的环境变量。
- (2) 设置容器启动命令的启动参数（需设置为环境变量）。
- (3) 以 Volume 的形式挂载为容器内部的文件或目录。

ConfigMap 以一个或多个 key:value 的形式保存在 Kubernetes 系统中供应用使用，既可以用于表示一个变量的值（例如 apploglevel=info），也可以用于表示一个完整配置文件的内容（例如 server.xml=<?xml...>...）

可以通过 yaml 配置文件或者直接使用 kubectl create configmap 命令行的方式来创建 ConfigMap。

## 2. 创建 ConfigMap 资源对象

### 1) 通过 yaml 配置文件方式创建

下面的例子 cm-appvars.yaml 描述了将几个应用所需的变量定义为 ConfigMap 的用法:

```
cm-appvars.yaml
apiVersion: v1
kind: ConfigMap
metadata:
  name: cm-appvars
data:
  apploglevel: info
  appdatadir: /var/data
```

执行 kubectl create 命令创建该 ConfigMap:

```
$kubectl create -f cm-appvars.yaml
configmap "cm-appvars" created
```

查看创建好的 ConfigMap:

```
# kubectl get configmap
NAME          DATA      AGE
cm-appvars    2          3s

# kubectl describe configmap cm-appvars
Name:         cm-appvars
Namespace:    default
Labels:       <none>
Annotations:  <none>

Data
====
appdatadir:   9 bytes
apploglevel:  4 bytes

# kubectl get configmap cm-appvars -o yaml
apiVersion: v1
data:
  appdatadir: /var/data
  apploglevel: info
kind: ConfigMap
metadata:
  creationTimestamp: 2016-07-28T19:57:16Z
  name: cm-appvars
  namespace: default
  resourceVersion: "78709"
  selfLink: /api/v1/namespaces/default/configmaps/cm-appvars
```

```
uid: 7bb2e9c0-54fd-11e6-9dcd-000c29dc2102
```

下面的例子 `cm-appconfigfiles.yaml` 描述了将两个配置文件 `server.xml` 和 `logging.properties` 定义为 `ConfigMap` 的用法，设置 `key` 为配置文件的别名，`value` 则是配置文件的全部文本内容：

```
cm-appconfigfiles.yaml
apiVersion: v1
kind: ConfigMap
metadata:
  name: cm-appconfigfiles
data:
  key-serverxml: |
    <?xml version='1.0' encoding='utf-8'?>
    <Server port="8005" shutdown="SHUTDOWN">
      <Listener className="org.apache.catalina.startup.VersionLoggerListener" />
      <Listener className="org.apache.catalina.core.AprLifecycleListener"
SSLEngine="on" />
      <Listener className=
"org.apache.catalina.core.JreMemoryLeakPreventionListener" />
      <Listener className=
"org.apache.catalina.mbeans.GlobalResourcesLifecycleListener" />
      <Listener className=
"org.apache.catalina.core.ThreadLocalLeakPreventionListener" />
      <GlobalNamingResources>
        <Resource name="UserDatabase" auth="Container"
          type="org.apache.catalina.UserDatabase"
          description="User database that can be updated and saved"
          factory="org.apache.catalina.users.MemoryUserDatabaseFactory"
          pathname="conf/tomcat-users.xml" />
      </GlobalNamingResources>

    <Service name="Catalina">
      <Connector port="8080" protocol="HTTP/1.1"
        connectionTimeout="20000"
        redirectPort="8443" />
      <Connector port="8009" protocol="AJP/1.3" redirectPort="8443" />
      <Engine name="Catalina" defaultHost="localhost">
        <Realm className="org.apache.catalina.realm.LockOutRealm">
          <Realm className="org.apache.catalina.realm.UserDatabaseRealm"
            resourceName="UserDatabase"/>
        </Realm>
        <Host name="localhost" appBase="webapps"
          unpackWARs="true" autoDeploy="true">
          <Valve className="org.apache.catalina.valves.AccessLogValve"
directory="logs"
            prefix="localhost_access_log" suffix=".txt"
            pattern="%h %l %u %t &quot;%r&quot; %s %b" />
```

```

        </Host>
    </Engine>
</Service>
</Server>
key-loggingproperties: "handlers
    =1catalina.org.apache.juli.FileHandler, 2localhost.org.apache.juli.
FileHandler,
    3manager.org.apache.juli.FileHandler, 4host-manager.org.apache.juli.
FileHandler,
    java.util.logging.ConsoleHandler\r\n\r\n.handlers= 1catalina.org.apache.
juli.FileHandler,

java.util.logging.ConsoleHandler\r\n\r\n1catalina.org.apache.juli.FileHandler.level
    = FINE\r\n1catalina.org.apache.juli.FileHandler.directory =
${catalina.base}/logs\r\n1catalina.org.apache.juli.FileHandler.prefix
    = catalina.\r\n\r\n2localhost.org.apache.juli.FileHandler.level =
FINE\r\n2localhost.org.apache.juli.FileHandler.directory
    = ${catalina.base}/logs\r\n2localhost.org.apache.juli.FileHandler.prefix =
localhost.\r\n\r\n3manager.org.apache.juli.FileHandler.level
    = FINE\r\n3manager.org.apache.juli.FileHandler.directory =
${catalina.base}/logs\r\n3manager.org.apache.juli.FileHandler.prefix
    = manager.\r\n\r\n4host-manager.org.apache.juli.FileHandler.level =
FINE\r\n4host-manager.org.apache.juli.FileHandler.directory
    = ${catalina.base}/logs\r\n4host-manager.org.apache.juli.FileHandler.
prefix =
    host-manager.\r\n\r\njava.util.logging.ConsoleHandler.level = FINE\r\n
njava.util.logging.ConsoleHandler.formatter
    = java.util.logging.SimpleFormatter\r\n\r\n\r\norg.apache.catalina.core.
ContainerBase.[Catalina].[localhost].level
    = INFO\r\norg.apache.catalina.core.ContainerBase.[Catalina].[localhost].
handlers
    = 2localhost.org.apache.juli.FileHandler\r\n\r\norg.apache.catalina.core.
ContainerBase.[Catalina].[localhost].[/manager].level
    = INFO\r\norg.apache.catalina.core.ContainerBase.[Catalina].[localhost].
[/manager].handlers
    = 3manager.org.apache.juli.FileHandler\r\n\r\norg.apache.catalina.core.
ContainerBase.[Catalina].[localhost].[/host-manager].level
    = INFO\r\norg.apache.catalina.core.ContainerBase.[Catalina].[localhost].
[/host-manager].handlers
    = 4host-manager.org.apache.juli.FileHandler\r\n\r\n"

```

执行 **kubect**l create 命令创建该 ConfigMap:

```
$kubectl create -f cm-appconfigfiles.yaml
configmap "cm-appconfigfiles" created
```

查看创建好的 ConfigMap:

```
# kubectl get configmap cm-appconfigfiles
NAME          DATA      AGE
cm-appconfigfiles  2          14s

# kubectl describe configmap cm-appconfigfiles
Name:          cm-appconfigfiles
Namespace:     default
Labels:        <none>
Annotations:   <none>
```

#### Data

====

**key-loggingproperties: 1809 bytes**

**key-serverxml: 1686 bytes**

查看已创建的 ConfigMap 的详细内容，可以看到两个配置文件的全文：

```
# kubectl get configmap cm-appconfigfiles -o yaml
apiVersion: v1
data:
  key-loggingproperties: "handlers = 1catalina.org.apache.juli.FileHandler,
2localhost.org.apache.juli.FileHandler,
3manager.org.apache.juli.FileHandler, 4host-manager.org.apache.juli.
FileHandler,
java.util.logging.ConsoleHandler\r\n\r\nhandlers = 1catalina.org.apache.
juli.FileHandler,
java.util.logging.ConsoleHandler\r\n\r\n1catalina.org.apache.juli.
FileHandler.level
= FINE\r\n1catalina.org.apache.juli.FileHandler.directory =
${catalina.base}/logs\r\n1catalina.org.apache.juli.FileHandler.prefix
= catalina.\r\n\r\n2localhost.org.apache.juli.FileHandler.level =
FINE\r\n2localhost.org.apache.juli.FileHandler.directory
= ${catalina.base}/logs\r\n2localhost.org.apache.juli.FileHandler.prefix =
localhost.\r\n\r\n3manager.org.apache.juli.FileHandler.level
= FINE\r\n3manager.org.apache.juli.FileHandler.directory =
${catalina.base}/logs\r\n3manager.org.apache.juli.FileHandler.prefix
= manager.\r\n\r\n4host-manager.org.apache.juli.FileHandler.level =
FINE\r\n4host-manager.org.apache.juli.FileHandler.directory
= ${catalina.base}/logs\r\n4host-manager.org.apache.juli.FileHandler.
prefix =
host-manager.\r\n\r\njava.util.logging.ConsoleHandler.level = FINE\r\njava.
util.logging.ConsoleHandler.formatter
= java.util.logging.SimpleFormatter\r\n\r\n\r\n\r\norg.apache.catalina.core.
ContainerBase.[Catalina].[localhost].level
= INFO\r\norg.apache.catalina.core.ContainerBase.[Catalina].[localhost].
handlers
= 2localhost.org.apache.juli.FileHandler\r\n\r\n\r\norg.apache.catalina.core.
ContainerBase.[Catalina].[localhost].[/manager].level
```



```

        = INFO\r\norg.apache.catalina.core.ContainerBase.[Catalina].[localhost].
[/manager].handlers
        = 3manager.org.apache.juli.FileHandler\r\n\r\norg.apache.catalina.core.
ContainerBase.[Catalina].[localhost].[/host-manager].level
        = INFO\r\norg.apache.catalina.core.ContainerBase.[Catalina].[localhost].
[/host-manager].handlers
        = 4host-manager.org.apache.juli.FileHandler\r\n\r\n"
key-serverxml: |
    <?xml version='1.0' encoding='utf-8'?>
    <Server port="8005" shutdown="SHUTDOWN">
        <Listener className="org.apache.catalina.startup.VersionLoggerListener" />
        <Listener className="org.apache.catalina.core.AprLifecycleListener"
SSLEngine="on" />
        <Listener className="org.apache.catalina.core.
JreMemoryLeakPreventionListener" />
        <Listener className="org.apache.catalina.mbeans.
GlobalResourcesLifecycleListener" />
        <Listener className="org.apache.catalina.core.
ThreadLocalLeakPreventionListener" />
        <GlobalNamingResources>
            <Resource name="UserDatabase" auth="Container"
                type="org.apache.catalina.UserDatabase"
                description="User database that can be updated and saved"
                factory="org.apache.catalina.users.MemoryUserDatabaseFactory"
                pathname="conf/tomcat-users.xml" />
        </GlobalNamingResources>

    <Service name="Catalina">
        <Connector port="8080" protocol="HTTP/1.1"
            connectionTimeout="20000"
            redirectPort="8443" />
        <Connector port="8009" protocol="AJP/1.3" redirectPort="8443" />
        <Engine name="Catalina" defaultHost="localhost">
            <Realm className="org.apache.catalina.realm.LockOutRealm">
                <Realm className="org.apache.catalina.realm.UserDatabaseRealm"
                    resourceName="UserDatabase"/>
            </Realm>
            <Host name="localhost" appBase="webapps"
                unpackWARs="true" autoDeploy="true">
                <Valve className="org.apache.catalina.valves.AccessLogValve"
directory="logs"
                    prefix="localhost_access_log" suffix=".txt"
                    pattern="%h %l %u %t &quot;%r&quot; %s %b" />
            </Host>
        </Engine>
    </Service>

```

```
</Server>
kind: ConfigMap
metadata:
  creationTimestamp: 2016-07-29T00:52:18Z
  name: cm-appconfigfiles
  namespace: default
  resourceVersion: "85054"
  selfLink: /api/v1/namespaces/default/configmaps/cm-appconfigfiles
  uid: b30d5019-5526-11e6-9dcd-000c29dc2102
```

## 2) 通过 kubectl 命令行方式创建

不使用 yaml 文件，直接通过 `kubectl create configmap` 也可以创建 ConfigMap，可以使用参数 `--from-file` 或 `--from-literal` 指定内容，并且可以在一行命令中指定多个参数。

(1) 通过 `--from-file` 参数从文件中进行创建，可以指定 key 的名称，也可以在一个命令行中创建包含多个 key 的 ConfigMap，语法为：

```
# kubectl create configmap NAME --from-file=[key=]source --from-file=[key=]source
```

(2) 通过 `--from-file` 参数从目录中进行创建，该目录下的每个配置文件名都被设置为 key，文件的内容被设置为 value，语法为：

```
# kubectl create configmap NAME --from-file=config-files-dir
```

(3) `--from-literal` 从文本中进行创建，直接将指定的 `key#=value#` 创建为 ConfigMap 的内容，语法为：

```
# kubectl create configmap NAME --from-literal=key1=value1 --from-literal=key2=value2
```

下面对这几种用法举例说明。

例如，当前目录下含有配置文件 `server.xml`，可以创建一个包含该文件内容的 ConfigMap：

```
# kubectl create configmap cm-server.xml --from-file=server.xml
configmap "cm-server.xml" created
```

```
# kubectl describe configmap cm-server.xml
Name:          cm-server.xml
Namespace:     default
Labels:        <none>
Annotations:   <none>
```

```
Data
====
server.xml:    6458 bytes
```

假设 `configfiles` 目录下包含两个配置文件 `server.xml` 和 `logging.properties`，创建一个包含这两个文件内容的 ConfigMap：

```
# kubectl create configmap cm-appconf --from-file=configfiles
configmap "cm-appconf" created
```

```
# kubectl describe configmap cm-appconf
Name:          cm-appconf
Namespace:     default
Labels:        <none>
Annotations:   <none>
```

Data

====

```
logging.properties: 3354 bytes
server.xml:         6458 bytes
```

使用--from-literal 参数进行创建的示例如下：

```
# kubectl create configmap cm-appenv --from-literal=loglevel=info --from-literal
=appdatadir=/var/data
configmap "cm-appenv" created
```

```
# kubectl describe configmap cm-appenv
Name:          cm-appenv
Namespace:     default
Labels:        <none>
Annotations:   <none>
```

Data

====

```
appdatadir: 9 bytes
loglevel:   4 bytes
```

容器应用对 ConfigMap 的使用有以下两种方法。

- (1) 通过环境变量获取 ConfigMap 中的内容。
- (2) 通过 Volume 挂载的方式将 ConfigMap 中的内容挂载为容器内部的文件或目录。

### 3. 在 Pod 中使用 ConfigMap

#### 1) 通过环境变量方式使用 ConfigMap

以前面创建的 ConfigMap “cm-appvars” 为例：

```
apiVersion: v1
kind: ConfigMap
metadata:
  name: cm-appvars
data:
  apploglevel: info
```

```
appdatadir: /var/data
```

在 Pod “cm-test-pod” 的定义中，将 ConfigMap “cm-appvars” 中的内容以环境变量（APPLOGLEVEL 和 APPDATADIR）设置为容器内部的环境变量，容器的启动命令将显示这两个环境变量的值（“env | grep APP”）：

```
apiVersion: v1
kind: Pod
metadata:
  name: cm-test-pod
spec:
  containers:
  - name: cm-test
    image: busybox
    command: [ "/bin/sh", "-c", "env | grep APP" ]
    env:
      - name: APPLOGLEVEL          # 定义环境变量名称
        valueFrom:                 # key “apploglevel” 对应的值
          configMapKeyRef:
            name: cm-appvars       # 环境变量的值取自 cm-appvars 中：
            key: apploglevel       # key 为 “apploglevel”
      - name: APPDATADIR          # 定义环境变量名称
        valueFrom:                 # key “appdatadir” 对应的值
          configMapKeyRef:
            name: cm-appvars       # 环境变量的值取自 cm-appvars 中：
            key: appdatadir        # key 为 “appdatadir”
    restartPolicy: Never
```

使用 `kubect1 create -f` 命令创建该 Pod，由于是测试 Pod，所以该 Pod 在执行完启动命令后将会退出，并且不会被系统自动重启（`restartPolicy=Never`）：

```
# kubect1 create -f cm-test-pod.yaml
pod "cm-test-pod" created
```

使用 `kubect1 get pods --show-all` 查看已经停止的 Pod：

```
# kubect1 get pods --show-all
NAME          READY    STATUS      RESTARTS   AGE
cm-test-pod   0/1      Completed   0           8s
```

查看该 Pod 的日志，可以看到启动命令 “env | grep APP” 的执行结果如下：

```
# kubect1 logs cm-test-pod
APPDATADIR=/var/data
APPLOGLEVEL=info
```

说明容器内部的环境变量使用 ConfigMap cm-appvars 中的值进行了正确的设置。

从 Kubernetes v1.6 开始，引入了一个新的字段 `envFrom`，实现在 Pod 环境内将 ConfigMap（也可用于 Secret 资源对象）中所有定义的 `key=value` 自动生成成为环境变量：

```

apiVersion: v1
kind: Pod
metadata:
  name: cm-test-pod
spec:
  containers:
  - name: cm-test
    image: busybox
    command: [ "/bin/sh", "-c", "env" ]
    envFrom:
    - configMapRef
      name: cm-appvars      # 根据 cm-appvars 中的 key=value 自动生成环境变量
  restartPolicy: Never

```

通过这个定义，在容器内部将会生成如下环境变量：

```

apploglevel=info
appdatadir=/var/data

```

需要说明的是，环境变量的名称受 POSIX 命名规范（[a-zA-Z\_][a-zA-Z0-9\_]\*）约束，不能以数字开头。如果包含非法字符，则系统将跳过该条环境变量的创建，并记录一个 Event 来描述环境变量无法生成，但并不阻止 Pod 的启动。

## 2) 通过 volumeMount 使用 ConfigMap

下面所示的 cm-appconfigfiles.yaml 例子中包含两个配置文件的定义：server.xml 和 logging.properties。

```

cm-appconfigfiles.yaml
apiVersion: v1
kind: ConfigMap
metadata:
  name: cm-serverxml
data:
  key-serverxml: |
    <?xml version='1.0' encoding='utf-8'?>
    <Server port="8005" shutdown="SHUTDOWN">
      <Listener className="org.apache.catalina.startup.VersionLoggerListener" />
      <Listener className="org.apache.catalina.core.AprLifecycleListener"
        SSLEngine="on" />
      <Listener className="org.apache.catalina.core.
        JreMemoryLeakPreventionListener" />
      <Listener className="org.apache.catalina.mbeans.
        GlobalResourcesLifecycleListener" />
      <Listener className="org.apache.catalina.core.
        ThreadLocalLeakPreventionListener" />
      <GlobalNamingResources>
        <Resource name="UserDatabase" auth="Container"

```

```

        type="org.apache.catalina.UserDatabase"
        description="User database that can be updated and saved"
        factory="org.apache.catalina.users.MemoryUserDatabaseFactory"
        pathname="conf/tomcat-users.xml" />
</GlobalNamingResources>

<Service name="Catalina">
    <Connector port="8080" protocol="HTTP/1.1"
        connectionTimeout="20000"
        redirectPort="8443" />
    <Connector port="8009" protocol="AJP/1.3" redirectPort="8443" />
    <Engine name="Catalina" defaultHost="localhost">
        <Realm className="org.apache.catalina.realm.LockOutRealm">
            <Realm className="org.apache.catalina.realm.UserDatabaseRealm"
                resourceName="UserDatabase"/>
        </Realm>
        <Host name="localhost" appBase="webapps"
            unpackWARs="true" autoDeploy="true">
            <Valve className="org.apache.catalina.valves.AccessLogValve"
directory="logs"
                prefix="localhost_access_log" suffix=".txt"
                pattern="%h %l %u %t &quot;%r&quot; %s %b" />
        </Host>
    </Engine>
</Service>
</Server>
key-loggingproperties: "handlers
    = 1catalina.org.apache.juli.FileHandler,
2localhost.org.apache.juli.FileHandler,
    3manager.org.apache.juli.FileHandler,
4host-manager.org.apache.juli.FileHandler,
    java.util.logging.ConsoleHandler\r\n\r\n.handlers =
1catalina.org.apache.juli.FileHandler,

java.util.logging.ConsoleHandler\r\n\r\n1catalina.org.apache.juli.FileHandler.level
    = FINE\r\n1catalina.org.apache.juli.FileHandler.directory =
${catalina.base}/logs\r\n1catalina.org.apache.juli.FileHandler.prefix
    = catalina.\r\n\r\n2localhost.org.apache.juli.FileHandler.level =
FINE\r\n2localhost.org.apache.juli.FileHandler.directory
    = ${catalina.base}/logs\r\n2localhost.org.apache.juli.FileHandler.prefix =
localhost.\r\n\r\n3manager.org.apache.juli.FileHandler.level
    = FINE\r\n3manager.org.apache.juli.FileHandler.directory =
${catalina.base}/logs\r\n3manager.org.apache.juli.FileHandler.prefix
    = manager.\r\n\r\n4host-manager.org.apache.juli.FileHandler.level =
FINE\r\n4host-manager.org.apache.juli.FileHandler.directory
    = ${catalina.base}/logs\r\n4host-manager.org.apache.juli.FileHandler.

```

```

prefix =
    host-manager.\r\n\r\njava.util.logging.ConsoleHandler.level =
FINE\r\njava.util.logging.ConsoleHandler.formatter
    = java.util.logging.SimpleFormatter\r\n\r\n\r\norg.apache.catalina.core.
ContainerBase.[Catalina].[localhost].level
    = INFO\r\norg.apache.catalina.core.ContainerBase.[Catalina].[localhost].
handlers
    = 2localhost.org.apache.juli.FileHandler\r\n\r\n\r\norg.apache.catalina.core.
ContainerBase.[Catalina].[localhost].[/manager].level
    = INFO\r\norg.apache.catalina.core.ContainerBase.[Catalina].[localhost].
[/manager].handlers
    = 3manager.org.apache.juli.FileHandler\r\n\r\n\r\norg.apache.catalina.core.
ContainerBase.[Catalina].[localhost].[/host-manager].level
    = INFO\r\norg.apache.catalina.core.ContainerBase.[Catalina].[localhost].
[/host-manager].handlers
    = 4host-manager.org.apache.juli.FileHandler\r\n\r\n\r\n

```

在 Pod “cm-test-app” 的定义中，将 ConfigMap “cm-appconfigfiles” 中的内容以文件的形式 mount 到容器内部的/configfiles 目录中去。Pod 配置文件 cm-test-app.yaml 的内容如下：

```

apiVersion: v1
kind: Pod
metadata:
  name: cm-test-app
spec:
  containers:
  - name: cm-test-app
    image: kubeguide/tomcat-app:v1
    ports:
    - containerPort: 8080
    volumeMounts:
    - name: serverxml          # 引用 volume 名
      mountPath: /configfiles  # 挂载到容器内的目录
  volumes:
  - name: serverxml          # 定义 volume 名
    configMap:
      name: cm-appconfigfiles # 使用 ConfigMap “cm-appconfigfiles”
      items:
      - key: key-serverxml     # key=key-serverxml
        path: server.xml       # value 将 server.xml 文件名进行挂载
      - key: key-loggingproperties # key=key-loggingproperties
        path: logging.properties # value 将 logging.properties 文件名进行挂载

```

创建该 Pod：

```

# kubectl create -f cm-test-app.yaml
pod "cm-test-app" created

```

登录容器，查看到/configfiles 目录下存在 server.xml 和 logging.properties 文件，它们的内容

就是 ConfigMap “cm-appconfigfiles” 中两个 key 定义的内容。

```
# kubectl exec -ti cm-test-app -- bash
root@cm-test-app:/# cat /configfiles/server.xml
<?xml version='1.0' encoding='utf-8'?>
<Server port="8005" shutdown="SHUTDOWN">
.....

root@cm-test-app:/# cat /configfiles/logging.properties
handlers = 1catalina.org.apache.juli.AsyncFileHandler,
2localhost.org.apache.juli.AsyncFileHandler,
3manager.org.apache.juli.AsyncFileHandler,
4host-manager.org.apache.juli.AsyncFileHandler, java.util.logging.ConsoleHandler
.....
```

如果在引用 ConfigMap 时不指定 items，则使用 volumeMount 方式在容器内的目录中为每个 item 生成一个文件名为 key 的文件。

Pod 配置文件 cm-test-app.yaml 的内容如下：

```
apiVersion: v1
kind: Pod
metadata:
  name: cm-test-app
spec:
  containers:
  - name: cm-test-app
    image: kubeguide/tomcat-app:v1
    imagePullPolicy: Never
    ports:
    - containerPort: 8080
    volumeMounts:
    - name: serverxml          # 引用 volume 名
      mountPath: /configfiles # 挂载到容器内的目录
  volumes:
  - name: serverxml          # 定义 volume 名
    configMap:
      name: cm-appconfigfiles # 使用 ConfigMap “cm-appconfigfiles”
```

创建该 Pod：

```
# kubectl create -f cm-test-app.yaml
pod "cm-test-app" created
```

登录容器，查看到/configfiles 目录下存在 key-loggingproperties 和 key-serverxml 文件，文件的名称来自 ConfigMap cm-appconfigfiles 中定义的两个 key 的名称，文件的内容则为 value 的内容：

```
# ls /configfiles
```



```
key-loggingproperties key-server.xml
```

#### 4. 使用 ConfigMap 的限制条件

使用 ConfigMap 的限制条件如下。

- ◎ ConfigMap 必须在 Pod 之前创建。
- ◎ ConfigMap 受 Namespace 限制，只有处于相同 Namespaces 中的 Pod 可以引用它。
- ◎ ConfigMap 中的配额管理还未能实现。
- ◎ kubelet 只支持可以被 API Server 管理的 Pod 使用 ConfigMap。kubelet 在本 Node 上通过 --manifest-url 或 --config 自动创建的静态 Pod 将无法引用 ConfigMap。
- ◎ 在 Pod 对 ConfigMap 进行挂载（volumeMount）操作时，容器内部只能挂载为“目录”，无法挂载为“文件”。在挂载到容器内部后，目录中将包含 ConfigMap 定义的每个 item，如果该目录下原来还有其他文件，则容器内的该目录将会被挂载的 ConfigMap 覆盖。如果应用程序需要保留原来的其他文件，则需要进行额外的处理。可以将 ConfigMap 挂载到容器内部的临时目录，再通过启动脚本将配置文件复制或者链接到（cp 或 link 命令）应用所用的实际配置目录下。

### 2.3.6 在容器内获取 Pod 信息（Downward API）

我们知道，每个 Pod 在成功创建出来之后，都会被系统分配唯一的名字、IP 地址，并且处于某个 Namespace 中，那么我们如何在 Pod 的容器内获取 Pod 的这些重要信息呢？答案就是使用 Downward API。

Downward API 可以通过以下两种方式将 Pod 信息注入容器内部。

- （1）环境变量：用于单个变量，可以将 Pod 信息和 Container 信息注入容器内部。
- （2）Volume 挂载：将数组类信息生成为文件，挂载到容器内部。

下面通过几个例子对 Downward API 的用法进行说明。

#### 例 1：环境变量方式——将 Pod 信息注入为环境变量

下例通过 Downward API 将 Pod 的 IP、名称和所在 Namespace 注入容器的环境变量中，容器应用使用 env 命令将全部环境变量打印到标准输出中：

```
dapi-test-pod.yaml
apiVersion: v1
kind: Pod
```

```
metadata:
  name: dapi-test-pod
spec:
  containers:
    - name: test-container
      image: busybox
      command: [ "/bin/sh", "-c", "env" ]
      env:
        - name: MY_POD_NAME
          valueFrom:
            fieldRef:
              fieldPath: metadata.name
        - name: MY_POD_NAMESPACE
          valueFrom:
            fieldRef:
              fieldPath: metadata.namespace
        - name: MY_POD_IP
          valueFrom:
            fieldRef:
              fieldPath: status.podIP
      restartPolicy: Never
```

注意到上面 valueFrom 这种特殊的语法是 Downward API 的写法。目前 Downward API 提供了以下变量。

- ◎ metadata.name: Pod 的名称，当 Pod 通过 RC 生成时，其名称是 RC 随机产生的唯一名称。
- ◎ status.podIP: Pod 的 IP 地址，之所以叫作 status.podIP 而非 metadata.IP，是因为 Pod 的 IP 属于状态数据，而非元数据。
- ◎ metadata.namespace: Pod 所在的 Namespace。

运行 kubectl create 命令创建 Pod:

```
# kubectl create -f dapi-test-pod.yaml
pod "dapi-test-pod" created
```

查看 dapi-test-pod 的日志:

```
# kubectl logs dapi-test-pod
.....
MY_POD_NAMESPACE=default
MY_POD_IP=172.17.1.2
MY_POD_NAME=dapi-test-pod
.....
```

从日志中我们可以看到 Pod 的 IP、Name 及 Namespace 等信息都被正确保存到了 Pod 的环境变量中。

**例 2：环境变量方式：将容器资源信息注入为环境变量**

下例通过 Downward API 将 Container 的资源请求和限制信息注入容器的环境变量中，容器应用使用 `printenv` 命令将设置的资源请求和资源限制环境变量打印到标准输出中：

```
dapi-test-pod-container-vars.yaml
apiVersion: v1
kind: Pod
metadata:
  name: dapi-test-pod-container-vars
spec:
  containers:
    - name: test-container
      image: busybox
      imagePullPolicy: Never
      command: [ "sh", "-c" ]
      args:
        - while true; do
          echo -en '\n';
          printenv MY_CPU_REQUEST MY_CPU_LIMIT;
          printenv MY_MEM_REQUEST MY_MEM_LIMIT;
          sleep 3600;
        done;
      resources:
        requests:
          memory: "32Mi"
          cpu: "125m"
        limits:
          memory: "64Mi"
          cpu: "250m"
      env:
        - name: MY_CPU_REQUEST
          valueFrom:
            resourceFieldRef:
              containerName: test-container
              resource: requests.cpu
        - name: MY_CPU_LIMIT
          valueFrom:
            resourceFieldRef:
              containerName: test-container
              resource: limits.cpu
        - name: MY_MEM_REQUEST
          valueFrom:
            resourceFieldRef:
              containerName: test-container
              resource: requests.memory
        - name: MY_MEM_LIMIT
```

```
    valueFrom:
      resourceFieldRef:
        containerName: test-container
        resource: limits.memory
  restartPolicy: Never
```

注意 valueFrom 这种特殊的 Downward API 语法，目前 resourceFieldRef 可以将容器的资源请求和资源限制等配置设置为容器内部的环境变量。

- ⊙ requests.cpu：容器的 CPU 请求值。
- ⊙ limits.cpu：容器的 CPU 限制值。
- ⊙ requests.memory：容器的内存请求值。
- ⊙ limits.memory：容器的内存限制值。

运行 `kubectl create` 命令来创建 Pod：

```
# kubectl create -f dapi-test-pod-container-vars.yaml
pod "dapi-test-pod-container-vars" created

# kubectl get pods
NAME                                READY   STATUS    RESTARTS   AGE
dapi-test-pod-container-vars       1/1     Running   0           36s
```

查看 `dapi-test-pod-container-vars` 的日志：

```
# kubectl logs dapi-test-pod-container-vars
1
1
33554432
67108864
```

从日志中我们可以看到 Container 的 requests.cpu、limits.cpu、requests.memory、limits.memory 等信息都被正确保存到了 Pod 的环境变量中。

### 例 3：Volume 挂载方式

下例通过 Downward API 将 Pod 的 Label、Annotation 列表通过 Volume 挂载为容器内的一个文件，容器应用使用 `echo` 命令将文件内容打印到标准输出中：

```
dapi-test-pod-volume.yaml
apiVersion: v1
kind: Pod
metadata:
  name: dapi-test-pod-volume
  labels:
    zone: us-est-coast
```

```

cluster: test-cluster1
rack: rack-22
annotations:
  build: two
  builder: john-doe
spec:
  containers:
    - name: test-container
      image: busybox
      imagePullPolicy: Never
      command: ["sh", "-c"]
      args:
        - while true; do
            if [[ -e /etc/labels ]]; then
              echo -en '\n\n'; cat /etc/labels; fi;
            if [[ -e /etc/annotations ]]; then
              echo -en '\n\n'; cat /etc/annotations; fi;
            sleep 3600;
          done;
      volumeMounts:
        - name: podinfo
          mountPath: /etc
          readOnly: false
  volumes:
    - name: podinfo
      downwardAPI:
        items:
          - path: "labels"
            fieldRef:
              fieldPath: metadata.labels
          - path: "annotations"
            fieldRef:
              fieldPath: metadata.annotations

```

注意 volumes 中 downwardAPI 的特殊语法，通过 items 的设置，将会以 path 的名称生成文件。这里将在容器内生成 /etc/labels 和 /etc/annotations 两个文件，/etc/labels 中将包含 metadata.labels 的全部 Label 列表，/etc/annotations 中将包含 metadata.annotations 的全部 Label 列表。

运行 kubectl create 命令创建 Pod:

```

# kubectl create -f dapi-test-pod-volume.yaml
pod "dapi-test-pod-volume" created

# kubectl get pods

```

NAME	READY	STATUS	RESTARTS	AGE
dapi-test-pod-volume	1/1	Running	0	1m

查看 dapi-test-pod-volume 的日志：

```
# k logs dapi-test-pod-volume
cluster="test-cluster1"
rack="rack-22"
zone="us-est-coast"

build="two"
builder="john-doe"
```

从日志中我们看到 Pod 的 Label 和 Annotation 信息都被保存到了容器内的/etc/labels 和 /etc/annotations 文件中。

那么，Downward API 有什么价值呢？

在某些集群中，集群中的每个节点都需要将自身的标识（ID）及进程绑定的 IP 地址等信息事先写入配置文件中，进程启动时读取这些信息，然后发布到某个类似服务注册中心的地方，以实现集群节点的自动发现功能。此时 Downward API 就可以派上用场了，具体做法是先编写一个预启动脚本或 Init Container，通过环境变量或文件方式获取 Pod 自身的名称、IP 地址等信息，然后写入主程序的配置文件中，最后启动主程序。

### 2.3.7 Pod 生命周期和重启策略

Pod 在整个生命周期过程中被系统定义为各种状态，熟悉 Pod 的各种状态对于我们理解如何设置 Pod 的调度策略、重启策略是很有必要的。

Pod 的状态如表 2.14 所示。

表 2.14 Pod 的状态

状态值	描述
Pending	API Server 已经创建该 Pod，但 Pod 内还有一个或多个容器的镜像没有创建，包括正在下载镜像的过程
Running	Pod 内所有容器均已创建，且至少有一个容器处于运行状态、正在启动状态或正在重启状态
Succeeded	Pod 内所有容器均成功执行退出，且不会再重启
Failed	Pod 内所有容器均已退出，但至少有一个容器退出为失败状态
Unknown	由于某种原因无法获取该 Pod 的状态，可能由于网络通信不畅导致

Pod 的重启策略（RestartPolicy）应用于 Pod 内的所有容器，并且仅在 Pod 所处的 Node 上由 kubelet 进行判断和重启操作。当某个容器异常退出或者健康检查（详见下节）失败时，kubelet 将根据 RestartPolicy 的设置来进行相应的操作。

Pod 的重启策略包括 Always、OnFailure 和 Never，默认值为 Always。

- Always：当容器失效时，由 kubelet 自动重启该容器。

- ◎ **OnFailure:** 当容器终止运行且退出码不为 0 时，由 kubelet 自动重启该容器。
- ◎ **Never:** 不论容器运行状态如何，kubelet 都不会重启该容器。

kubelet 重启失效容器的时间间隔以 `sync-frequency` 乘以  $2n$  来计算，例如 1、2、4、8 倍等，最长延时 5min，并且在成功重启后的 10min 后重置该时间。

Pod 的重启策略与控制方式息息相关，当前可用于管理 Pod 的控制器包括 ReplicationController、Job、DaemonSet 及直接通过 kubelet 管理（静态 Pod）。每种控制器对 Pod 的重启策略要求如下。

- ◎ **RC 和 DaemonSet:** 必须设置为 Always，需要保证该容器持续运行。
- ◎ **Job:** OnFailure 或 Never，确保容器执行完成后不再重启。
- ◎ **kubelet:** 在 Pod 失效时自动重启它，不论将 RestartPolicy 设置为什么值，也不会对 Pod 进行健康检查。

结合 Pod 的状态和重启策略，表 2.15 列出一些常见的状态转换场景。

表 2.15 一些常见的状态转换场景

Pod 包含的容器数	Pod 当前的状态	发 生 事 件	Pod 的结果状态		
			RestartPolicy= Always	RestartPolicy= OnFailure	RestartPolicy= Never
包含 1 个容器	Running	容器成功退出	Running	Succeeded	Succeeded
包含 1 个容器	Running	容器失败退出	Running	Running	Failed
包含两个容器	Running	1 个容器失败退出	Running	Running	Running
包含两个容器	Running	容器被 OOM 杀掉	Running	Running	Failed

### 2.3.8 Pod 健康检查

对 Pod 的健康状态检查可以通过两类探针来检查：LivenessProbe 和 ReadinessProbe。

- ◎ **LivenessProbe 探针:** 用于判断容器是否存活（running 状态），如果 LivenessProbe 探针探测到容器不健康，则 kubelet 将杀掉该容器，并根据容器的重启策略做相应的处理。如果一个容器不包含 LivenessProbe 探针，那么 kubelet 认为该容器的 LivenessProbe 探针返回的值永远是“Success”。
- ◎ **ReadinessProbe 探针:** 用于判断容器是否启动完成（ready 状态），可以接收请求。如果 ReadinessProbe 探针检测到失败，则 Pod 的状态将被修改。Endpoint Controller 将从 Service 的 Endpoint 中删除包含该容器所在 Pod 的 Endpoint。

kubelet 定期执行 LivenessProbe 探针来诊断容器的健康状况。LivenessProbe 有以下三种实现方式。

(1) **ExecAction**: 在容器内部执行一个命令，如果该命令的返回码为 0，则表明容器健康。

在下面的例子中，通过执行“cat /tmp/health”命令来判断一个容器运行是否正常。而该 Pod 运行之后，在创建/tmp/health 文件的 10s 之后将删除该文件，而 LivenessProbe 健康检查的初始探测时间（initialDelaySeconds）为 15s，探测结果将是 Fail，将导致 kubelet 杀掉该容器并重启它。

```
apiVersion: v1
kind: Pod
metadata:
  labels:
    test: liveness
  name: liveness-exec
spec:
  containers:
  - name: liveness
    image: gcr.io/google_containers/busybox
    args:
    - /bin/sh
    - -c
    - echo ok > /tmp/health; sleep 10; rm -rf /tmp/health; sleep 600
    livenessProbe:
      exec:
        command:
        - cat
        - /tmp/health
      initialDelaySeconds: 15
      timeoutSeconds: 1
```

(2) **TCPSocketAction**: 通过容器的 IP 地址和端口号执行 TCP 检查，如果能够建立 TCP 连接，则表明容器健康。

在下面的例子中，通过与容器内的 localhost:80 建立 TCP 连接进行健康检查。

```
apiVersion: v1
kind: Pod
metadata:
  name: pod-with-healthcheck
spec:
  containers:
  - name: nginx
    image: nginx
    ports:
    - containerPort: 80
```



```

livenessProbe:
  tcpSocket:
    port: 80
  initialDelaySeconds: 30
  timeoutSeconds: 1

```

(3) HTTPGetAction: 通过容器的 IP 地址、端口号及路径调用 HTTP Get 方法, 如果响应的状态码大于等于 200 且小于 400, 则认为容器状态健康。

在下面的例子中, kubelet 定时发送 HTTP 请求到 localhost:80/\_status/healthz 来进行容器应用的健康检查。

```

apiVersion: v1
kind: Pod
metadata:
  name: pod-with-healthcheck
spec:
  containers:
  - name: nginx
    image: nginx
    ports:
    - containerPort: 80
    livenessProbe:
      httpGet:
        path: /_status/healthz
        port: 80
      initialDelaySeconds: 30
      timeoutSeconds: 1

```

对于每种探测方式, 都需要设置 initialDelaySeconds 和 timeoutSeconds 两个参数, 它们的含义分别如下。

- ◎ **initialDelaySeconds:** 启动容器后进行首次健康检查的等待时间, 单位为 s。
- ◎ **timeoutSeconds:** 健康检查发送请求后等待响应的超时时间, 单位为 s。当超时发生时, kubelet 会认为容器已经无法提供服务, 将会重启该容器。

### 2.3.9 玩转 Pod 调度

在 Kubernetes 系统中, Pod 在大部分场景下都只是容器的载体而已, 通常需要通过 Deployment、DaemonSet、RC、Job 等对象来完成一组 Pod 的调度与自动控制功能。

#### 1. Deployment/RC: 全自动调度

Deployment 或 RC 的主要功能之一就是自动部署一个容器应用的多份副本, 以及持续监控

副本的数量，在集群内始终维持用户指定的副本数量。

下面是一个 **Deployment** 配置的例子，使用这个配置文件可以创建一个 **ReplicaSet**，这个 **ReplicaSet** 会创建 3 个 Nginx 应用的 Pod。

```
nginx-deployment.yaml
apiVersion: apps/v1beta1
kind: Deployment
metadata:
  name: nginx-deployment
spec:
  replicas: 3
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx:1.7.9
        ports:
        - containerPort: 80
```

运行 **kubectl create** 命令创建这个 **Deployment**：

```
# kubectl create -f nginx-deployment.yaml
deployment "nginx-deployment" created
```

查看 **Deployment** 的状态：

```
# kubectl get deployments
NAME                DESIRED   CURRENT   UP-TO-DATE   AVAILABLE   AGE
nginx-deployment    3         3         3            3           18s
```

该状态说明 **Deployment** 已创建好所有 3 个副本，并且所有副本都是最新的而且是可用的。

运行 **kubectl get rs** 和 **kubectl get pods** 可以查看已创建的 **ReplicaSet**（RS）和 Pod 的信息。

```
# kubectl get rs
NAME                                DESIRED   CURRENT   READY   AGE
nginx-deployment-4087004473        3         3         3       53s

# kubectl get pods
NAME                                READY     STATUS    RESTARTS   AGE
nginx-deployment-4087004473-9jqqs  1/1       Running   0          1m
nginx-deployment-4087004473-cq0cf  1/1       Running   0          1m
nginx-deployment-4087004473-vxn56  1/1       Running   0          1m
```

从调度策略上来说，这 3 个 Nginx Pod 由系统全自动完成调度。它们各自最终运行在哪个节点上，完全由 Master 的 Scheduler 经过一系列算法计算得出，用户无法干预调度过程和结果。

除了使用系统自动调度算法完成一组 Pod 的部署，Kubernetes 也提供了多种丰富的调度策略，用户只需在 Pod 的定义中使用 NodeSelector、NodeAffinity、PodAffinity、Pod 驱逐等更加细粒度的调度策略设置，就能完成对 Pod 的精准调度。下面对这些策略进行说明。

## 2. NodeSelector: 定向调度

Kubernetes Master 上的 Scheduler 服务（kube-scheduler 进程）负责实现 Pod 的调度，整个调度过程通过执行一系列复杂的算法，最终为每个 Pod 计算出一个最佳的目标节点，这一过程是自动完成的，通常我们无法知道 Pod 最终会被调度到哪个节点上。在实际情况中，也可能需要将 Pod 调度到指定的一些 Node 上，可以通过 Node 的标签（Label）和 Pod 的 nodeSelector 属性相匹配，来达到上述目的。

(1) 首先通过 `kubectl label` 命令给目标 Node 打上一些标签：

```
kubectl label nodes <node-name> <label-key>=<label-value>
```

这里，我们为 `k8s-node-1` 节点打上一个 `zone=north` 的标签，表明它是“北方”的一个节点：

```
$ kubectl label nodes k8s-node-1 zone=north
```

NAME	LABELS	STATUS
k8s-node-1	kubernetes.io/hostname=k8s-node-1, <b>zone=north</b>	Ready

上述命令行操作也可以通过修改资源定义文件的方式，并执行 `kubectl replace -f xxx.yaml` 命令来完成。

(2) 然后，在 Pod 的定义中加上 nodeSelector 的设置，以 `redis-master-controller.yaml` 为例：

```
apiVersion: v1
kind: ReplicationController
metadata:
  name: redis-master
  labels:
    name: redis-master
spec:
  replicas: 1
  selector:
    name: redis-master
  template:
    metadata:
      labels:
        name: redis-master
    spec:
      containers:
        - name: master
          image: kubeguide/redis-master
          ports:
```

```
- containerPort: 6379
nodeSelector:
  zone: north
```

运行 `kubectl create -f` 命令创建 Pod，scheduler 就会将该 Pod 调度到拥有 `zone=north` 标签的 Node 上。

使用 `kubectl get pods -o wide` 命令可以验证 Pod 所在的 Node：

```
# kubectl get pods -o wide
NAME                READY   STATUS    RESTARTS   AGE      NODE
redis-master-f0rqj  1/1     Running   0           19s      k8s-node-1
```

如果我们给多个 Node 都定义了相同的标签（例如 `zone=north`），则 scheduler 将会根据调度算法从这组 Node 中挑选一个可用的 Node 进行 Pod 调度。

通过基于 Node 标签的调度方式，我们可以把集群中具有不同特点的 Node 贴上不同的标签，例如“`role=frontend`”“`role=backend`”“`role=database`”等标签，在部署应用时就可以根据应用的需求设置 NodeSelector 来进行指定 Node 范围的调度。

需要注意的是，如果我们指定了 Pod 的 nodeSelector 条件，且集群中不存在包含相应标签的 Node，则即使集群中还有其他可供使用的 Node，这个 Pod 也无法被成功调度。

除了用户可以自行给 Node 添加标签，Kubernetes 也会给 Node 预定义一些标签，包括：

- ◎ `kubernetes.io/hostname`
- ◎ `failure-domain.beta.kubernetes.io/zone`
- ◎ `failure-domain.beta.kubernetes.io/region`
- ◎ `beta.kubernetes.io/instance-type`
- ◎ `beta.kubernetes.io/os`
- ◎ `beta.kubernetes.io/arch`

用户也可以使用这些系统标签进行 Pod 的定向调度。

NodeSelector 通过标签的方式，简单地实现了限制 Pod 所在节点的方法。亲和性调度机制则极大地扩展了 Pod 的调度能力，主要的增强功能如下。

- ◎ 更具表达力（不仅仅是“符合全部”的简单情况）。
- ◎ 可以使用软限制、优先采用等限制方式，代替之前的硬限制，这样调度器在无法满足优先需求的情况下，会退而求其次，继续运行该 Pod。
- ◎ 可以依据节点上正在运行的其他 Pod 的标签来进行限制，而非节点本身的标签。这样就可以定义一种规则来描述 Pod 之间的亲和或互斥关系。

亲和性调度功能包括节点亲和性（NodeAffinity）和 Pod 亲和性（PodAffinity）两个维度的设置。节点亲和性与 NodeSelector 类似，增强了上述前两点的优势；Pod 的亲和与互斥限制则通过 Pod 标签而不是节点标签来实现，也就是上面第 4 点内容所陈述的方式，同时具有前两点提到的优点。

NodeSelector 将会继续使用，随着节点亲和性越来越能够表达 nodeSelector 表达的功能，最终 NodeSelector 会被废弃掉。

### 3. NodeAffinity: Node 亲和性调度

NodeAffinity 意为 Node 亲和性的调度策略，是用于替换 NodeSelector 的全新调度策略。目前有两种节点亲和性表达。

- ◎ RequiredDuringSchedulingIgnoredDuringExecution: 必须满足指定的规则才可以调度 Pod 到 Node 上（功能与 nodeSelector 很像，但是使用的是不同的语法），相当于硬限制。
- ◎ PreferredDuringSchedulingIgnoredDuringExecution: 强调优先满足指定规则，调度器会尝试调度 Pod 到 Node 上，但并不强求，相当于软限制。多个优先级规则还可以设置权重（weight）值，以定义执行的先后顺序。

IgnoredDuringExecution 的意思是：如果一个 Pod 所在的节点在 Pod 运行期间标签发生了变更，不再符合该 Pod 的节点亲和性需求，则系统将忽略 Node 上 Label 的变化，该 Pod 能继续在该节点运行。

下面的例子设置了 NodeAffinity 调度的如下规则。

- ◎ requiredDuringSchedulingIgnoredDuringExecution 要求只运行在 amd64 的节点上（beta.kubernetes.io/arch In amd64）。
- ◎ preferredDuringSchedulingIgnoredDuringExecution 的要求是尽量运行在“磁盘类型为 ssd”（disk-type In ssd）的节点上。

```
apiVersion: v1
kind: Pod
metadata:
  name: with-node-affinity
spec:
  affinity:
    nodeAffinity:
      requiredDuringSchedulingIgnoredDuringExecution:
        nodeSelectorTerms:
          - matchExpressions:
              - key: beta.kubernetes.io/arch
                operator: In
```

```
      values:
        - amd64
    preferredDuringSchedulingIgnoredDuringExecution:
    - weight: 1
      preference:
        matchExpressions:
        - key: disk-type
          operator: In
          values:
            - ssd
    containers:
    - name: with-node-affinity
      image: gcr.io/google_containers/pause:2.0
```

从上面的配置中可以看到 In 操作符，NodeAffinity 语法支持的操作符包括 In、NotIn、Exists、DoesNotExist、Gt、Lt。虽然没有节点排斥的功能，但是用 NotIn 和 DoesNotExist 就可以实现排斥的功能了。

NodeAffinity 规则设置的注意事项如下。

- ◎ 如果同时定义了 nodeSelector 和 nodeAffinity，那么必须两个条件都得到满足，Pod 才能最终运行在指定的 Node 上。
- ◎ 如果 nodeAffinity 指定了多个 nodeSelectorTerms，那么只需要其中一个能够匹配成功即可。
- ◎ 如果 nodeSelectorTerms 中有多个 matchExpressions，则一个节点必须满足所有 matchExpressions 才能运行该 Pod。

#### 4. PodAffinity：Pod 亲和与互斥调度策略

Pod 间的亲和与互斥从 Kubernetes v1.4 版本开始引入。这一功能让用户从另外一个角度来限制 Pod 所能运行的节点：根据节点上正在运行的 Pod 的标签而不是节点的标签进行判断和调度，要求对节点和 Pod 两个条件进行匹配。这种规则可以描述为：如果在具有标签 X 的 Node 上运行了一个或者多个符合条件 Y 的 Pod，那么 Pod 应该（如果是互斥的情况，那么就变成拒绝）运行在这个 Node 上。

这里 X 指的是一个集群中的节点、机架、区域等概念，通过 Kubernetes 内置节点标签中的 key 来进行声明。这个 key 的名字为 topologyKey，意为表达节点所属的 topology 范围。

- ◎ kubernetes.io/hostname
- ◎ failure-domain.beta.kubernetes.io/zone
- ◎ failure-domain.beta.kubernetes.io/region

与节点不同的是，Pod 是属于某个命名空间的，所以条件 Y 表达的是一个或者全部命名空间中的一个 Label Selector。

和节点亲和相同，Pod 亲和与互斥的条件设置也是 `requiredDuringSchedulingIgnoredDuringExecution` 和 `preferredDuringSchedulingIgnoredDuringExecution`。Pod 亲和性定义于 PodSpec 的 `affinity` 字段下的 `podAffinity` 子字段里。Pod 间的互斥性则定义于同一层次的 `podAntiAffinity` 子字段中。

下面通过实例来说明 Pod 间的亲和性和互斥性策略设置。

### 1) 参照目标 Pod

首先，创建一个名为 `pod-flag` 的 Pod，带有标签 `security=S1` 和 `app=nginx`，后面的例子将使用 `pod-flag` 作为 Pod 亲和与互斥的目标 Pod。

```
apiVersion: v1
kind: Pod
metadata:
  name: pod-flag
  labels:
    security: "S1"
    app: "nginx"
spec:
  containers:
  - name: nginx
    image: nginx
```

### 2) Pod 的亲和性调度

下面创建第 2 个 Pod 来说明 Pod 的亲和性调度，这里定义的亲和标签是 `security=S1`，对应上面的 Pod “`pod-flag`”，`topologyKey` 的值被设置为 “`kubernetes.io/hostname`”：

```
apiVersion: v1
kind: Pod
metadata:
  name: pod-affinity
spec:
  affinity:
    podAffinity:
      requiredDuringSchedulingIgnoredDuringExecution:
      - labelSelector:
          matchExpressions:
          - key: security
            operator: In
            values:
            - S1
        topologyKey: kubernetes.io/hostname
  containers:
  - name: with-pod-affinity
    image: gcr.io/google_containers/pause:2.0
```

创建 Pod 之后，使用 `kubecttl get pods -o wide` 命令可以看到，这两个 Pod 处于同一个 Node 之上运行。

有兴趣的读者还可以测试一下，在创建这个 Pod 之前，删掉这个节点的 `kubernetes.io/hostname` 标签，重复上面的创建步骤，将会发现 Pod 会一直处于 Pending 状态，这是因为找不到满足条件的 Node 了。

### 3) Pod 的互斥性调度

创建第 3 个 Pod，我们希望它不能与参照目标 Pod 运行在同一个 Node 上。

```
apiVersion: v1
kind: Pod
metadata:
  name: anti-affinity
spec:
  affinity:
    podAffinity:
      requiredDuringSchedulingIgnoredDuringExecution:
        - labelSelector:
            matchExpressions:
              - key: security
                operator: In
                values:
                  - S1
          topologyKey: failure-domain.beta.kubernetes.io/zone
    podAntiAffinity:
      requiredDuringSchedulingIgnoredDuringExecution:
        - labelSelector:
            matchExpressions:
              - key: app
                operator: In
                values:
                  - nginx
          topologyKey: kubernetes.io/hostname
  containers:
    - name: anti-affinity
      image: gcr.io/google_containers/pause:2.0
```

这里要求这个新 Pod 与 `security=S1` 的 Pod 为同一个 zone，但是不与 `app=nginx` 的 Pod 为同一个 Node。创建 Pod 之后，同样用 `kubecttl get pods -o wide` 来查看，会看到新的 Pod 被调度到了同一 Zone 内的不同的 Node 上去。

与节点亲和性类似，Pod 亲和性的操作符也包括 In、NotIn、Exists、DoesNotExist、Gt、Lt。

原则上，`topologyKey` 可以使用任何合法的标签 Key 赋值，但是出于性能和安全方面的考虑，对 `topologyKey` 有如下限制。



- ◎ 在 Pod 亲和性和 RequiredDuringScheduling 的 Pod 互斥性的定义中，不允许使用空的 topologyKey。
- ◎ 如果 Admission controller 包含了 LimitPodHardAntiAffinityTopology，那么针对 Required DuringScheduling 的 Pod 互斥性定义就被限制为 kubernetes.io/hostname，要使用自定义的 topologyKey，就要改写或禁用该控制器。
- ◎ 在 PreferredDuringScheduling 类型的 Pod 互斥性定义中，空的 topologyKey 会被解释为 kubernetes.io/hostname、failure-domain.beta.kubernetes.io/zone 及 failure-domain.beta.kubernetes.io/region 的组合。
- ◎ 如果不是上述情况，就可以采用任意合法的 topologyKey 了。

PodAffinity 规则设置的注意事项如下。

- ◎ 除了设置 Label Selector 和 topologyKey，用户还可以指定 namespace 列表来进行限制，同样，使用 Label Selector 对 namespace 进行选择。namespace 的定义和 Label Selector 及 topologyKey 同级。省略 namespace 的设置，表示使用定义了 affinity/anti-affinity 的 Pod 所在的 namespace。如果 namespace 设置为空值（""），则表示所有 namespace。
- ◎ 在所有关联 requiredDuringSchedulingIgnoredDuringExecution 的 matchExpressions 全都满足之后，系统才能将 Pod 调度到某个 Node 上。

更多关于 Pod 亲和性和互斥性调度的信息可以参考其设计文档，网址为 <https://github.com/kubernetes/kubernetes/blob/master/docs/design/podaffinity.md>。

## 5. Taints 和 Tolerations（污点和容忍）

前面介绍的 NodeAffinity 节点亲和性，是在 Pod 上定义的一种属性，使得 Pod 能够被调度到某些 Node 上运行（优先选择或强制要求）。Taint 则正好相反——它让 Node 拒绝 Pod 的运行。

Taint 需要和 Toleration 配合使用，让 Pod 避开那些不合适的 Node。在 Node 上设置一个或多个 Taint 之后，除非 Pod 明确声明能够容忍这些“污点”，否则无法在这些 Node 上运行。Toleration 是 Pod 的属性，让 Pod 能够（注意，只是能够，而非必须）运行在标注了 Taint 的 Node 上。

可以用 `kubectt taint` 命令为 Node 设置 Taint 信息：

```
$ kubectt taint nodes node1 key=value:NoSchedule
```

这个设置为 node1 加上了一个 Taint。该 Taint 的键为 key，值为 value，Taint 的效果是 NoSchedule。这意味着除非 Pod 明确声明可以容忍这个 Taint，否则就不会被调度到 node1 上去。

然后，需要在 Pod 上声明 Toleration。下面的两个 Toleration 都设置为可以容忍（Tolerate）具有该 Taint 的 Node，使得 Pod 能够被调度到 node1 上：

```
tolerations:
- key: "key"
  operator: "Equal"
  value: "value"
  effect: "NoSchedule"
```

或

```
tolerations:
- key: "key"
  operator: "Exists"
  effect: "NoSchedule"
```

Pod 的 Toleration 声明中的 key 和 effect 需要与 Taint 的设置保持一致，并且满足以下条件之一。

- ⊙ operator 的值是 Exists（无须指定 value）。
- ⊙ operator 的值是 Equal 并且 value 相等。

如果不指定 operator，则默认值为 Equal。

另外，还有如下两个特例。

- ⊙ 空的 key 配合 Exists 操作符能够匹配所有的键和值。
- ⊙ 空的 effect 匹配所有的 effect。

上面的例子中 effect 的取值为 NoSchedule，还可以取值为 PreferNoSchedule，这个值的意思是优先，也可以算作 NoSchedule 的软限制版本——一个 Pod 如果没有声明容忍这个 Taint，则系统会尽量避免把这个 Pod 调度到这一节点上去，但不是强制的。后面还会介绍另一个 effect “NoExecute”。

系统允许在同一个 Node 上设置多个 Taint，也可以在 Pod 上设置多个 Toleration。Kubernetes 调度器处理多个 Taint 和 Toleration 的逻辑顺序为：首先列出节点中所有的 Taint，然后忽略 Pod 的 Toleration 能够匹配的部分，剩下的没有忽略掉的 Taint 就是对 Pod 的效果了。下面是几种特殊情况。

- ⊙ 如果剩余的 Taint 中存在 effect=NoSchedule，则调度器不会把该 Pod 调度到这一节点上。
- ⊙ 如果剩余 Taint 中没有 NoSchedule 效果，但是有 PreferNoSchedule 效果，则调度器会尝试不把这个 Pod 指派给这个节点。
- ⊙ 如果剩余 Taint 的效果有 NoExecute 的，并且这个 Pod 已经在该节点上运行，则会被驱逐；如果没有在该节点上运行，也不会再被调度到该节点上。

例如，我们这样对一个节点进行 Taint 设置：

```
$ kubectl taint nodes node1 key1=value1:NoSchedule
$ kubectl taint nodes node1 key1=value1:NoExecute
$ kubectl taint nodes node1 key2=value2:NoSchedule
```

然后在 Pod 上设置两个 Toleration：

```
tolerations:
- key: "key1"
  operator: "Equal"
  value: "value1"
  effect: "NoSchedule"
- key: "key1"
  operator: "Equal"
  value: "value1"
  effect: "NoExecute"
```

这样的结果是该 Pod 无法被调度到 node1 上去，这是因为第 3 个 Taint 没有匹配的 Toleration。但是如果该 Pod 已经在 node1 上运行了，那么在运行时设置上第 3 个 Taint，它还能继续在 node1 上运行，这是因为 Pod 可以容忍前两个 Taint。

一般来说，如果给 Node 加上 effect=NoExecute 的 Taint，那么该 Node 上正在运行的所有无对应 Toleration 的 Pod 都会被立刻驱逐，而具有相应 Toleration 的 Pod 则永远不会被逐出。不过，系统允许给具有 NoExecute 效果的 Toleration 加入一个可选的 tolerationSeconds 字段，这个设置表明 Pod 可以在 Taint 添加到 Node 之后还能在这个 Node 上运行多久（单位为 s）：

```
tolerations:
- key: "key1"
  operator: "Equal"
  value: "value1"
  effect: "NoExecute"
  tolerationSeconds: 3600
```

上述定义的意思是，如果 Pod 正在运行，所在节点被加入一个匹配的 Taint，则这个 Pod 会持续在这个节点上存活 3600s 后被逐出。如果在这个宽限期内，Taint 被移除，则不会触发驱逐事件。

Taint 和 Toleration 是一种处理节点并且让 Pod 进行规避或者驱逐 Pod 的弹性处理方式，下面列举一些常见的用例。

### 1) 独占节点

如果想要拿出一部分节点，专门给一些特定应用使用，则可以为节点添加这样的 Taint：

```
$ kubectl taint nodes nodename dedicated=groupName:NoSchedule
```

然后给这些应用的 Pod 加入对应的 Toleration。这样，带有合适 Toleration 的 Pod 就会被允

许同使用其他节点一样使用有 Taint 的节点。

通过自定义 Admission Controller 也可以实现这一目标。如果希望让这些应用独占一批节点，并且确保它们只能使用这些节点，则还可以给这些 Taint 节点加入类似的标签（Label）`dedicated=groupName`，然后 Admission Controller 需要加入节点亲和性设置，要求 Pod 只会被调度到具有这一标签的节点上。

## 2) 具有特殊硬件设备的节点

在集群里可能有一小部分节点安装了特殊的硬件设备（如 GPU 芯片），用户自然会希望把不需要占用这类硬件的 Pod 排除在外，以确保对这类硬件有需求的 Pod 能够顺利调度到这些节点上。

可以用下面的命令为节点设置 Taint：

```
$ kubectl taint nodes nodename special=true:NoSchedule
$ kubectl taint nodes nodename special=true:PreferNoSchedule
```

然后在 Pod 中利用对应的 Toleration 来保障特定的 Pod 能够使用特定的硬件。

和上面的独占节点的示例类似，使用 Admission Controller 来完成这一任务会更方便。例如 Admission Controller 使用 Pod 的一些特征来判断这些 Pod，如果可以使用这些硬件，就添加 Toleration 来完成这一工作。要保障需要使用特殊硬件的 Pod 只被调度到安装这些硬件的节点上，则还需要一些额外的工作，比如将这些特殊资源使用 `opaque-int-resource` 的方式对自定义资源进行量化，然后在 PodSpec 中进行请求；也可以使用标签的方式来标注这些安装有特别硬件的节点，然后在 Pod 中定义节点亲和性来实现这个目标。

## 3) 定义 Pod 驱逐行为，以应对节点故障（为 Alpha 版本的功能）

前面提到的 NoExecute 这个 Taint 效果对节点上正在运行的 Pod 有以下影响。

- ◎ 没有设置 Toleration 的 Pod 会被立刻驱逐。
- ◎ 配置了对应 Toleration 的 Pod，如果没有为 `tolerationSeconds` 赋值，则会一直留在这一节点中。
- ◎ 配置了对应 Toleration 的 Pod 且指定了 `tolerationSeconds` 值，则会在指定时间后驱逐。
- ◎ 从 Kubernetes v1.6 版本开始引入一个 Alpha 版本的功能，即把节点故障标记为 Taint（目前只针对 `node unreachable` 及 `node not ready`，相应的 NodeCondition "Ready" 的值分别为 `Unknown` 和 `False`）。激活 `TaintBasedEvictions` 功能后（在 `--feature-gates` 参数中加入 `TaintBasedEvictions=true`），NodeController 会自动为 Node 设置 Taint，而状态为“Ready”的 Node 上之前设置过的普通驱逐逻辑将会被禁用。注意，在节点故障情况下，为了保持现存的 Pod 驱逐的限速（rate-limiting）设置，系统将会以限速的模式逐步给 Node

设置 Taint，这就能防止在一些特定情况下（比如 Master 暂时失联）造成的大量 Pod 被驱逐的后果。这一功能兼容于 `tolerationSeconds`，允许 Pod 定义节点故障时持续多久才被逐出。

例如一个包含很多本地状态的应用可能需要在网络发生故障时，还能持续在节点上运行，期望网络能够快速恢复，从而避免从这个 Node 上被驱逐。

Pod 的 Toleration 可以这样定义：

```
tolerations:
- key: "node.alpha.kubernetes.io/unreachable"
  operator: "Exists"
  effect: "NoExecute"
  tolerationSeconds: 6000
```

对于 Node 未就绪状态，可以把 Key 设置为 `node.alpha.kubernetes.io/notReady`。

如果没有为 Pod 指定 `node.alpha.kubernetes.io/notReady` 的 Toleration，那么 Kubernetes 会自动为 Pod 加入 `tolerationSeconds=300` 的 `node.alpha.kubernetes.io/notReady` 类型的 Toleration。

同样，如果 Pod 没有定义 `node.alpha.kubernetes.io/unreachable` 的 Toleration，那么系统会自动为其加入 `tolerationSeconds=300` 的 `node.alpha.kubernetes.io/unreachable` 类型的 Toleration。

这些系统自动设置的 toleration 用于在 Node 发现问题时，能够为 Pod 确保驱逐前再运行 5min。这两个默认的 Toleration 由 Admission Controller “DefaultTolerationSeconds” 自动加入。

## 6. DaemonSet：在每个 Node 上调度一个 Pod

DemonSet 是 Kubernetes v1.2 版本新增的一种资源对象，用于管理在集群中每个 Node 上仅运行一份 Pod 的副本实例，如图 2.4 所示。

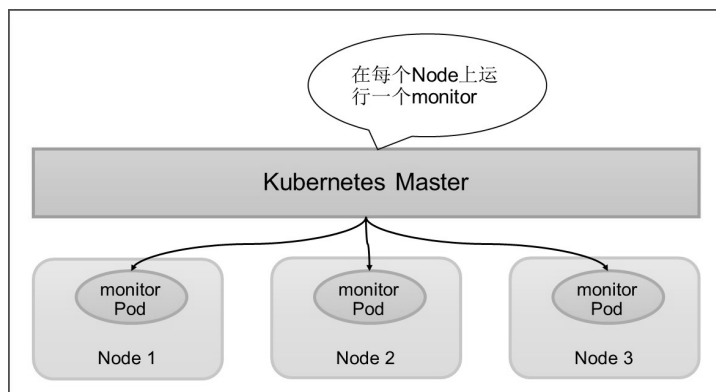


图 2.4 DaemonSet 示例

这种用法适合一些有这种需求的应用。

- ◎ 在每个 Node 上运行一个 GlusterFS 存储或者 Ceph 存储的 Daemon 进程。
- ◎ 在每个 Node 上运行一个日志采集程序，例如 Fluentd 或者 Logstash。
- ◎ 在每个 Node 上运行一个性能监控程序，采集该 Node 的运行性能数据，例如 Prometheus Node Exporter、collectd、New Relic agent 或者 Ganglia gmond 等。

DaemonSet 的 Pod 调度策略与 RC 类似，除了使用系统内置的算法在每台 Node 上进行调度，也可以在 Pod 的定义中使用 NodeSelector 或 NodeAffinity 来指定满足条件的 Node 范围进行调度。

下面的例子定义为在每台 Node 上启动一个 fluentd 容器，配置文件 fluentd-ds.yaml 的内容如下，其中挂载了物理机的两个目录 “/var/log” 和 “/var/lib/docker/containers”：

```
apiVersion: extensions/v1beta1
kind: DaemonSet
metadata:
  name: fluentd-cloud-logging
  namespace: kube-system
  labels:
    k8s-app: fluentd-cloud-logging
spec:
  template:
    metadata:
      namespace: kube-system
      labels:
        k8s-app: fluentd-cloud-logging
    spec:
      containers:
      - name: fluentd-cloud-logging
        image: gcr.io/google_containers/fluentd-elasticsearch:1.17
        resources:
          limits:
            cpu: 100m
            memory: 200Mi
        env:
        - name: FLUENTD_ARGS
          value: -q
        volumeMounts:
        - name: varlog
          mountPath: /var/log
          readOnly: false
        - name: containers
          mountPath: /var/lib/docker/containers
          readOnly: false
```

```
volumes:
- name: containers
  hostPath:
    path: /var/lib/docker/containers
- name: varlog
  hostPath:
    path: /var/log
```

使用 `kubectl create` 命令创建该 `DaemonSet`:

```
# kubectl create -f fluentd-ds.yaml
daemonset "fluentd-cloud-logging" created
```

查看创建好的 `DaemonSet` 和 `Pod`，可以看到在每个 `Node` 上都创建了一个 `Pod`:

```
# kubectl get daemonset --namespace=kube-system
NAME                                DESIRED  CURRENT  NODE-SELECTOR  AGE
fluentd-cloud-logging              2        2        <none>         3s

# kubectl get pods --namespace=kube-system
NAME                                READY    STATUS    RESTARTS  AGE
fluentd-cloud-logging-7tw9z        1/1     Running   0          1h
fluentd-cloud-logging-aqdn1        1/1     Running   0          1h
```

## 7. Job: 批处理调度

Kubernetes 从 1.2 版本开始支持批处理类型的应用，我们可以通过 `Kubernetes Job` 资源对象来定义并启动一个批处理任务。批处理任务通常并行（或者串行）启动多个计算进程去处理一批工作项（`work item`），处理完成后，整个批处理任务结束。按照批处理任务实现方式的不同，批处理任务可以分为如图 2.5 所示的几种模式。

- ◎ **Job Template Expansion 模式**: 一个 `Job` 对象对应一个待处理的 `Work item`，有几个 `Work item` 就产生几个独立的 `Job`，通常适合 `Work item` 数量少、每个 `Work item` 要处理的数据量比较大的场景，比如有一个 100GB 的文件作为一个 `Work item`，总共 10 个文件需要处理。
- ◎ **Queue with Pod Per Work Item 模式**: 采用一个任务队列存放 `Work item`，一个 `Job` 对象作为消费者去完成这些 `Work item`，在这种模式下，`Job` 会启动  $N$  个 `Pod`，每个 `Pod` 对应一个 `Work item`。
- ◎ **Queue with Variable Pod Count 模式**: 也是采用一个任务队列存放 `Work item`，一个 `Job` 对象作为消费者去完成这些 `Work item`，但与上面的模式不同，`Job` 启动的 `Pod` 数量是可变的。

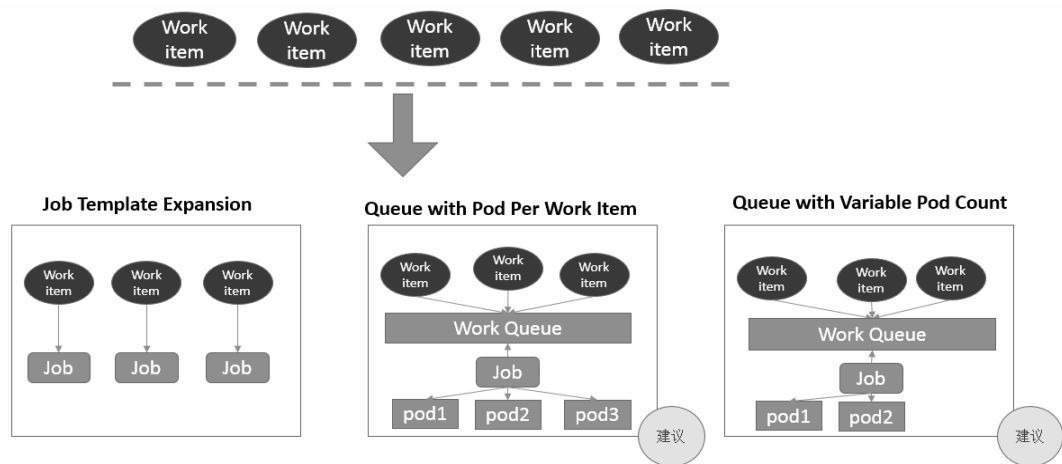


图 2.5 批处理任务的几种模式

还有一种被称为 Single Job with Static Work Assignment 的模式，也是一个 Job 产生多个 Pod 的模式，但它采用程序静态方式分配任务项，而不是采用队列模式进行动态分配。

如表 2.16 所示是这几种模式的一个对比。

表 2.16 批处理任务的模式对比

模式名称	是否是一个 Job	Pod 的数量少于 Work item	用户程序是否要做相应的修改	Kubernetes 是否支持
Job Template Expansion	/	/	是	是
Queue with Pod Per Work Item	是	/	有时候需要	是
Queue with Variable Pod Count	是	/	/	是
Single Job with Static Work Assignment	是	/	是	/

考虑到批处理的并行问题，Kubernetes 将 Job 分以下三种类型。

1) Non-parallel Jobs

通常一个 Job 只启动一个 Pod，则除非 Pod 异常，才会重启该 Pod，一旦此 Pod 正常结束，Job 将结束。

2) Parallel Jobs with a fixed completion count

并行 Job 会启动多个 Pod，此时需要设定 Job 的.spec.completions 参数为一个正数，当正常结束的 Pod 数量达至此参数设定的值后，Job 结束。此外，Job 的.spec.parallelism 参数用来控制并行度，即同时启动几个 Job 来处理 Work Item。



### 3) Parallel Jobs with a work queue

任务队列方式的并行 Job 需要一个独立的 Queue，Work item 都在一个 Queue 中存放，不能设置 Job 的.spec.completions 参数，此时 Job 有以下特性。

- ◎ 每个 Pod 能独立判断和决定是否还有任务项需要处理。
- ◎ 如果某个 Pod 正常结束，则 Job 不会再启动新的 Pod。
- ◎ 如果一个 Pod 成功结束，则此时应该不存在其他 Pod 还在干活的情况，它们应该都处于即将结束、退出的状态。
- ◎ 如果所有 Pod 都结束了，且至少有一个 Pod 成功结束，则整个 Job 算是成功结束。

下面我们分别说说常见的三种批处理模型在 Kubernetes 中的例子。

首先是 Job Template Expansion 模式，由于这种模式下每个 Work item 对应一个 Job 实例，所以这种模式首先定义一个 Job 模板，模板里主要的参数是 Work item 的标识，因为每个 Job 处理不同的 Work item。如下所示的 Job 模板（文件名为 job.yaml.txt）中的\$ITEM 可以作为任务项的标识：

```
apiVersion: batch/v1
kind: Job
metadata:
  name: process-item-$ITEM
  labels:
    jobgroup: jobexample
spec:
  template:
    metadata:
      name: jobexample
      labels:
        jobgroup: jobexample
    spec:
      containers:
        - name: c
          image: busybox
          command: ["sh", "-c", "echo Processing item $ITEM && sleep 5"]
          restartPolicy: Never
```

通过下面的操作，生成 3 个对应的 Job 定义文件并创建 Job：

```
# for i in apple banana cherry
> do
>   cat job.yaml.txt | sed "s/\$ITEM/\$i/" > ./jobs/job-$(i).yaml
> done
# ls jobs
job-apple.yaml  job-banana.yaml  job-cherry.yaml
```

```
# kubectl create -f jobs
job "process-item-apple" created
job "process-item-banana" created
job "process-item-cherry" created
```

首先，观察 Job 的运行情况：

```
# kubectl get jobs -l jobgroup=jobexample
NAME                DESIRED   SUCCESSFUL   AGE
process-item-apple   1         1            4m
process-item-banana  1         1            4m
process-item-cherry  1         1            4m
```

其次，我们看看 Queue with Pod Per Work Item 模式，在这种模式下需要一个任务队列存放 Work item，比如 RabbitMQ，客户端程序先把要处理的任务变成 Work item 放入到任务队列，然后编写 Worker 程序并打包镜像并定义成为 Job 中的 Work Pod，Worker 程序的实现逻辑是从任务队列中拉取一个 Work item 并处理，处理完成后即结束进程，图 2.6 给出了并行度为 2 的一个 Demo 示意图。

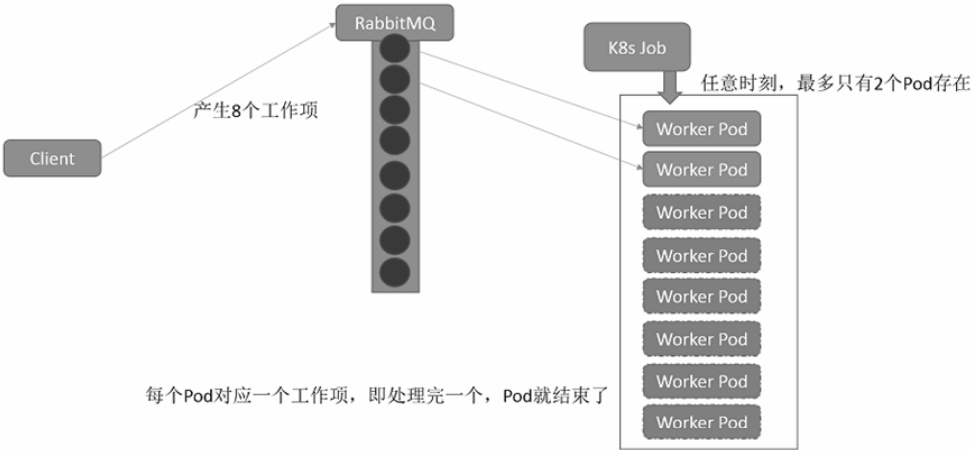


图 2.6 Queue with Pod Per Work Item 示例

最后，我们再看看 Queue with Variable Pod Count 模式，如图 2.7 所示，由于这种模式下，Worker 程序需要知道队列中是否还有等待处理的 Work item，如果有就取出来并处理，否则就认为所有工作完成并结束进程，所以任务队列通常要采用 Redis 或者数据库来实现。

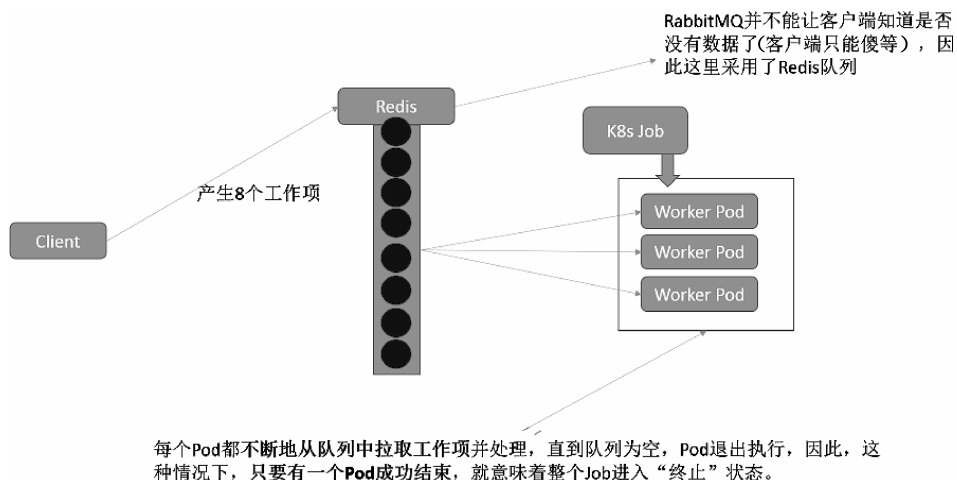


图 2.7 Queue with Variable Pod Count 示例

## 8. Cronjob: 定时任务

Kubernetes 从 v1.5 版本开始增加了一种新类型的 Job, 即类似 Linux Cron 的定时任务 Cron Job, 下面我们来看看如何定义和使用这种新类型的 Job。

首先, 确保 Kubernetes 的版本为 v1.5 及以上, 在 API Server 的启动进程上增加以下配置参数并重启:

```
--runtime-config=batch/v2alpha1=true
```

其次, 我们需要掌握 Cron Job 的定时表达式, 它基本上照搬了 Linux Cron 的表达式, 区别是第 1 位是分钟而不是秒, 格式如下:

```
Minutes Hours DayofMonth Month DayofWeek Year
```

其中每一个域可出现的字符如下。

- ⊙ **Minutes:** 可出现", - \* /"这 4 个字符, 有效范围为 0-59 的整数。
- ⊙ **Hours:** 可出现", - \* /"这 4 个字符, 有效范围为 0-23 的整数。
- ⊙ **DayofMonth:** 可出现", - \* / ? L W C"这 8 个字符, 有效范围为 0-31 的整数。
- ⊙ **Month:** 可出现", - \* /"这 4 个字符, 有效范围为 1-12 的整数或 JAN-DEC。
- ⊙ **DayofWeek:** 可出现", - \* / ? L C #"这 4 个字符, 有效范围为 1-7 的整数或 SUN-SAT 两个范围。1 表示星期天, 2 表示星期一, 以此类推。

表达式中的特殊字符 “\*” 与 “/” 的含义如下。

- ◎ \*：表示匹配该域的任意值，假如在 Minutes 域使用\*，则表示每分钟都会触发事件。
- ◎ /：表示起始时间开始触发，然后每隔固定时间触发一次，例如在 Minutes 域设置为 5/20，则意味着第 1 次触发在第 5min 时，接下来每 20min 触发一次，将在第 25min、第 45min 等时刻分别触发。

比如，我们要每隔 1min 执行一次任务，则 Cron 表达式如下：

```
*/1 * * * *
```

掌握这些基本知识后，就可以编写一个 Cron Job 的配置文件了：

```
cron.yaml
apiVersion: batch/v2alpha1
kind: CronJob
metadata:
  name: hello
spec:
  schedule: "*/1 * * * *"
  jobTemplate:
    spec:
      template:
        spec:
          containers:
            - name: hello
              image: busybox
              args:
                - /bin/sh
                - -c
                - date; echo Hello from the Kubernetes cluster
          restartPolicy: OnFailure
```

该例子定义了一个名为 hello 的 Cron Job，任务每隔 1min 执行一次，运行的镜像是 busybox，执行的命令是 shell 脚本，脚本执行时会在控制台输出当前时间和字符串“Hello from the Kubernetes cluster”。

接下来，执行 kubectl create 命令完成创建：

```
# kubectl create -f cron.yaml
cronjob "hello" created
```

然后，我们每隔 1min 执行 kubectl get cronjob hello 查看任务状态，发现的确是每分钟调度了一次：

```
# kubectl get cronjob hello
NAME          SCHEDULE          SUSPEND   ACTIVE   LAST-SCHEDULE
hello         */1 * * * *      False    0        Thu, 29 Jun 2017 11:32:00 -0700
.....
# kubectl get cronjob hello
```

```

NAME          SCHEDULE          SUSPEND   ACTIVE   LAST-SCHEDULE
hello         */1 * * * *       False    0        Thu, 29 Jun 2017 11:33:00 -0700
.....
# kubectl get cronjob hello
NAME          SCHEDULE          SUSPEND   ACTIVE   LAST-SCHEDULE
hello         */1 * * * *       False    0        Thu, 29 Jun 2017 11:34:00 -0700

```

还可以通过查找 Cron Job 对应的容器，验证每隔 1min 产生一个容器的事实，如下所示：

```

# docker ps -a | grep busybox
83f7b86728ea
busybox@sha256:be3c11fdb7cfe299214e46edc642e09514dbb9bbefcd0d3836c05a1e0cd0642
"/bin/sh -c 'date; ec"    About a minute ago    Exited (0) About a minute ago
k8s_hello_hello-1498795860-qgwb4_default_207586cf-5d4a-11e7-86c1-000c2997487d_0
36aa3b991980
busybox@sha256:be3c11fdb7cfe299214e46edc642e09514dbb9bbefcd0d3836c05a1e0cd0642
"/bin/sh -c 'date; ec"    2 minutes ago    Exited (0) 2 minutes ago
k8s_hello_hello-1498795800-g92vx_default_fca21ec0-5d49-11e7-86c1-000c2997487d_0
3d762ae35172
busybox@sha256:be3c11fdb7cfe299214e46edc642e09514dbb9bbefcd0d3836c05a1e0cd0642
"/bin/sh -c 'date; ec"    3 minutes ago    Exited (0) 3 minutes ago
k8s_hello_hello-1498795740-3qxmd_default_d8c75d07-5d49-11e7-86c1-000c2997487d_0
8ee5eefa8cd3
busybox@sha256:be3c11fdb7cfe299214e46edc642e09514dbb9bbefcd0d3836c05a1e0cd0642
"/bin/sh -c 'date; ec"    4 minutes ago    Exited (0) 4 minutes ago
k8s_hello_hello-1498795680-mgb7h_default_b4f7aec5-5d49-11e7-86c1-000c2997487d_0

```

查看任意一个容器的日志，结果如下：

```

# docker logs 83f7b86728ea
Thu Jun 29 18:33:07 UTC 2017
Hello from the Kubernetes cluster

```

运行下面的命令，可以让我们更加直观地了解 Cron Job 定期触发任务执行的历史和现状：

```

# kubectl get jobs --watch
NAME          DESIRED   SUCCESSFUL   AGE
hello-1498761060    1         1           31m
hello-1498761120    1         1           30m
hello-1498761180    1         1           29m
hello-1498761240    1         1           28m
hello-1498761300    1         1           27m
hello-1498761360    1         1           26m
hello-1498761420    1         1           25m

```

其中 SUCCESSFUL 列为 1 的每一行都是一个调度成功的 Job，以第 1 行的“hello-1498761060”的 Job 为例，它对应的 Pod 可以通过下面的方式得到：

```

# kubectl get pods --show-all | grep hello-1498761060
hello-1498761060-shpwx    0/1    Completed    0    39m

```

查看该 Pod 的日志：

```
# kubectl logs hello-1498761060-shpwx
Thu Jun 29 18:31:07 UTC 2017
Hello from the Kubernetes cluster
```

最后，当我们不需要某个 Cron Job 时，可以通过下面的命令删除它：

```
# kubectl delete cronjob hello
cronjob "hello" deleted
```

## 9. 自定义调度器

如果 Kubernetes 调度器的众多特性还无法满足我们的独特调度需求，则我们还可以用自己开发的调度器进行调度。从 v1.6 版本开始，Kubernetes 的多调度器特性也进入了快速发展阶段。

一般情况下，每个新 Pod 都会由默认的调度器进行调度。但是如果 Pod 中提供了自定义的调度器名称，那么默认的调度器就会忽略该 Pod，转由指定的调度器完成 Pod 的调度。

在下面的例子中，为 Pod 指定了一个名为 my-scheduler 的自定义调度器：

```
apiVersion: v1
kind: Pod
metadata:
  name: nginx
  labels:
    app: nginx
spec:
  schedulerName: my-scheduler
  containers:
  - name: nginx
    image: nginx
```

如果自定义的调度器还未在系统中部署，则默认的调度器会忽略这个 Pod，这个 Pod 将会永远处于 Pending 状态。

下面我们看看如何创建一个自定义的调度器。

可以用任何语言来实现简单或复杂的自定义调度器。下面的简单例子使用 Bash 脚本进行实现，调度策略为随机选择一个 Node（注意，这个调度器需要通过 kubectl proxy 来运行）。

```
#!/bin/bash
SERVER='localhost:8001'
while true;
do
    for PODNAME in $(kubectl --server $SERVER get pods -o json | jq '.items[] |
select(.spec.schedulerName == "my-scheduler") | select(.spec.nodeName == null)
| .metadata.name' | tr -d ' ');
    do
```

```

        NODES=$(kubectl --server $SERVER get nodes -o json | jq
'.items[].metadata.name' | tr -d ' ')
        NUMNODES=${#NODES[@]}
        CHOSEN=${NODES[$[ $RANDOM % $NUMNODES ]]}
        curl --header "Content-Type:application/json" --request POST --data
'{"apiVersion":"v1", "kind": "Binding", "metadata": {"name": "'$PODNAME'"}, "target":
{"apiVersion": "v1", "kind": "Node", "name":"' $CHOSEN' }}"'
http://$SERVER/api/v1/namespaces/default/pods/$PODNAME/binding/
        echo "Assigned $PODNAME to $CHOSEN"
    done
    sleep 1
done

```

一旦这个自定义调度器成功启动，则前面的 Pod 将会被正确调度到某个 Node 上。

### 2.3.10 Init Container（初始化容器）

在很多应用场景中，应用在启动之前都需要进行如下初始化操作。

- ◎ 等待其他关联组件正确运行（例如数据库或某个后台服务）。
- ◎ 基于环境变量或配置模板生成配置文件。
- ◎ 从远程数据库获取本地所需配置，或者将自身注册到某个中央数据库中。
- ◎ 下载相关依赖包，或者对系统进行一些预配置操作。

Kubernetes v1.3 引入了一个 Alpha 版本的新特性 init container（在 Kubernetes v1.5 时被更新为 Beta 版本），用于在启动应用容器（app container）之前启动一个或多个“初始化”容器，完成应用容器所需的预置条件，如图 2.8 所示。Init container 与应用容器本质上是一样的，但它们是仅运行一次就结束的任务，并且必须在成功执行完成后，系统才能继续执行下一个容器。根据 Pod 的重启策略（RestartPolicy），当 init container 执行失败，在设置了 RestartPolicy=Never 时，Pod 将会启动失败；而设置 RestartPolicy=Always 时，Pod 将会被系统自动重启。

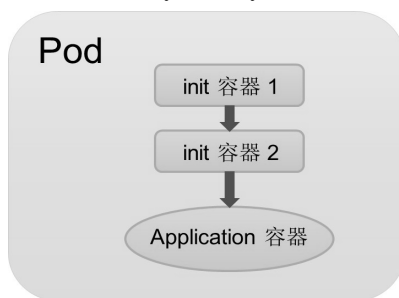


图 2.8 init container

下面以 Nginx 应用为例，在启动 Nginx 之前，通过初始化容器 busybox 为 Nginx 创建一个 index.html 主页文件。这里为 init container 和 Nginx 设置了一个共享的 volume，以供 Nginx 访问 init container 设置的 index.html 文件：

```
nginx-init-containers.yaml
apiVersion: v1
kind: Pod
metadata:
  name: nginx
  annotations:
spec:
  # These containers are run during pod initialization
  initContainers:
  - name: install
    image: busybox
    command:
      - wget
      - "-O"
      - "/work-dir/index.html"
      - "http://kubernetes.io"
    volumeMounts:
      - name: workdir
        mountPath: "/work-dir"
  containers:
  - name: nginx
    image: nginx
    ports:
      - containerPort: 80
    volumeMounts:
      - name: workdir
        mountPath: /usr/share/nginx/html
  dnsPolicy: Default
  volumes:
  - name: workdir
    emptyDir: {}
```

创建这个 Pod：

```
# kubectl create -f nginx-init-containers.yaml
pod "nginx" created
```

在运行 init container 的过程中，查看 Pod 的状态，可见 init 过程还未完成：

```
# kubectl get pods
NAME      READY   STATUS    RESTARTS   AGE
nginx     0/1     Init:0/1   0           1m
```

在 init container 成功执行完成后，系统继续启动 Nginx 容器，再次查看 Pod 的状态：



```
# kubectl get pods
NAME      READY   STATUS    RESTARTS   AGE
nginx     1/1     Running   0           7s
```

查看 Pod 的事件，可以看到系统首先创建并运行 init container 容器（名为 install），成功后继续创建和运行 Nginx 容器：

```
# kubectl describe pod nginx
Name:          nginx
Namespace:     default
..... (略)
Events:
  FirstSeen    LastSeen    Count   From          SubobjectPath
Type          Reason      Message
-----
-----
3s           3s           1       default-scheduler
Normal        Scheduled    Successfully assigned init-demo to k8s-node-1
3s           3s           1       kubelet, k8s-node-1
spec.initContainers{install} Normal        Pulled        Container image
"busybox" already present on machine
3s           3s           1       kubelet, k8s-node-1
spec.initContainers{install} Normal        Created        Created container
with id 93d98cbc0251c60d43c2d8d0a6a9bb65f432344fe6f04561c4a940b79bcff74a
3s           3s           1       kubelet, k8s-node-1
spec.initContainers{install} Normal        Started        Started container
with id 93d98cbc0251c60d43c2d8d0a6a9bb65f432344fe6f04561c4a940b79bcff74a
2s           2s           1       kubelet, k8s-node-1
spec.containers{nginx} Normal        Pulled        Container image
"nginx" already present on machine
2s           2s           1       kubelet, k8s-node-1
spec.containers{nginx} Normal        Created        Created container with
id a388bbb9f1fe247cf42e61449328ab20f7c54a7c271590548d3d8610a28a6048
1s           1s           1       kubelet, k8s-node-1
spec.containers{nginx} Normal        Started        Started container with
id a388bbb9f1fe247cf42e61449328ab20f7c54a7c271590548d3d8610a28a6048
```

启动成功后，登录进 Nginx 容器，可以查看到/usr/share/nginx/html 目录下的 index.html 文件为 init container 所生成，其内容为：

```
<html id="home" lang="en" class="">

<head>
...
<title>Kubernetes | Production-Grade Container Orchestration</title>
...
"url": "http://kubernetes.io/"</script>
</head>
```

```
<body>
...
```

init container 与应用容器的区别如下。

(1) init container 的运行方式与应用容器不同，它们必须先于应用容器执行完成，当设置了多个 init container 时，将按顺序逐个运行，并且只有前一个 init container 运行成功后才能运行后一个 init container。当所有 init container 都成功运行后，Kubernetes 才会初始化 Pod 的各种信息，并开始创建和运行应用容器。

(2) 在 init container 的定义中也可以设置资源限制、volume 的使用和安全策略，等等。但资源限制的设置与应用容器略有不同。

- ◎ 如果多个 init container 都定义了资源请求/资源限制，则取最大的值作为所有 init container 的资源请求值/资源限制值。
- ◎ Pod 的有效（effective）资源请求值/资源限制值取以下二者中的较大值。
  - a) 所有应用容器的资源请求值/资源限制值之和。
  - b) init container 的有效资源请求值/资源限制值。
- ◎ 调度算法将基于 Pod 的有效资源请求值/资源限制值进行计算，也就是说 init container 可以为初始化操作预留系统资源，即使后续应用容器无须使用这些资源。
- ◎ Pod 的有效 QoS 等级适用于 init container 和应用容器。
- ◎ 资源配额和限制将根据 Pod 的有效资源请求值/资源限制值计算生效。
- ◎ Pod 级别的 cgroup 将基于 Pod 的有效资源请求/限制，与调度机制一致。

(3) init container 不能设置 readinessProbe 探针，因为必须在它们成功运行后才能继续运行 Pod 中定义的普通容器。

在 Pod 重新启动（Restart）时，init container 将会重新运行，常见的 Pod 重启场景如下。

- ◎ init container 的镜像被更新时，init container 将会重新运行，导致 Pod 重启。仅更新应用容器的镜像只会使得应用容器被重启。
- ◎ Pod 的 infrastructure 容器（pause）更新时，Pod 将会重启。
- ◎ 若 Pod 中的所有应用容器都终止了，并且 RestartPolicy=Always，则 Pod 将会重启。

### 2.3.11 Pod 的升级和回滚

---

下面我们说说 Pod 的升级和回滚问题。

当集群中的某个服务需要升级时，我们需要停止目前与该服务相关的所有 Pod，然后下载新版本镜像并创建新的 Pod。如果集群规模比较大，则这个工作就变成了一个挑战，而且先全部停止然后逐步升级的方式会导致较长时间的服务不可用。Kubernetes 提供了滚动升级功能来解决上述问题。

如果 Pod 是通过 Deployment 创建的，则用户可以在运行时修改 Deployment 的 Pod 定义（spec.template）或镜像名称，并应用到 Deployment 对象上，系统即可完成 Deployment 的自动更新操作。如果在更新过程中发生了错误，则还可以通过回滚（Rollback）操作恢复 Pod 的版本。

## 1. Deployment 的升级

以 Deployment nginx 为例：

```
nginx-deployment.yaml
apiVersion: apps/v1beta1
kind: Deployment
metadata:
  name: nginx-deployment
spec:
  replicas: 3
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx:1.7.9
        ports:
        - containerPort: 80
```

已运行的 Pod 副本数量有 3 个：

```
# kubectl get pods
```

NAME	READY	STATUS	RESTARTS	AGE
nginx-deployment-4087004473-9jqqs	1/1	Running	0	1m
nginx-deployment-4087004473-cq0cf	1/1	Running	0	1m
nginx-deployment-4087004473-vxn56	1/1	Running	0	1m

现在 Pod 镜像需要更新为 Nginx:1.9.1，我们可以通过 kubectl set image 命令为 Deployment 设置新的镜像名称：

```
$ kubectl set image deployment/nginx-deployment nginx=nginx:1.9.1
deployment "nginx-deployment" image updated
```

另一种更新的方法是使用 `kubectl edit` 命令修改 Deployment 的配置，将 `spec.template.spec.containers[0].image` 从 `Nginx:1.7.9` 更改为 `Nginx:1.9.1`：

```
$ kubectl edit deployment/nginx-deployment
deployment "nginx-deployment" edited
```

一旦镜像名（或 Pod 定义）发生了修改，则将触发系统完成 Deployment 所有运行 Pod 的滚动升级操作。可以使用 `kubectl rollout status` 命令查看 Deployment 的更新过程：

```
$ kubectl rollout status deployment/nginx-deployment
Waiting for rollout to finish: 2 out of 3 new replicas have been updated...
Waiting for rollout to finish: 2 out of 3 new replicas have been updated...
Waiting for rollout to finish: 2 out of 3 new replicas have been updated...
Waiting for rollout to finish: 2 out of 3 new replicas have been updated...
Waiting for rollout to finish: 2 old replicas are pending termination...
Waiting for rollout to finish: 1 old replicas are pending termination...
Waiting for rollout to finish: 1 old replicas are pending termination...
Waiting for rollout to finish: 1 old replicas are pending termination...
Waiting for rollout to finish: 2 of 3 updated replicas are available...
deployment "nginx-deployment" successfully rolled out
```

查看当前运行的 Pod，名称已经更新了：

```
$ kubectl get pods
NAME                                READY    STATUS    RESTARTS   AGE
nginx-deployment-3599678771-01h26  1/1      Running   0           2m
nginx-deployment-3599678771-57thr  1/1      Running   0           2m
nginx-deployment-3599678771-s8p21  1/1      Running   0           2m
```

查看 Pod 使用的镜像，已经更新为 `Nginx:1.9.1` 了：

```
# kubectl describe pod/nginx-deployment-3599678771-s8p21
Name:                nginx-deployment-3599678771-s8p21
.....
    Image:            nginx:1.9.1
.....
```

那么，Deployment 是如何完成 Pod 更新的呢？

我们可以使用 `kubectl describe deployments/nginx-deployment` 命令仔细观察 Deployment 的更新过程。初始创建 Deployment 时，系统创建了一个 ReplicaSet (`nginx-deployment-4087004473`)，并按用户的需求创建了 3 个 Pod 副本。当更新 Deployment 时，系统创建了一个新的 ReplicaSet (`nginx-deployment-3599678771`)，并将其副本数扩展到 1，然后将旧的 ReplicaSet 缩减为 2。之后，系统继续按照相同的更新策略对新旧两个 ReplicaSet 进行逐个调整。最后，新的 ReplicaSet 运行了 3 个新版本 Pod 副本，旧的 ReplicaSet 副本数则缩减为 0。如图 2.9 所示。

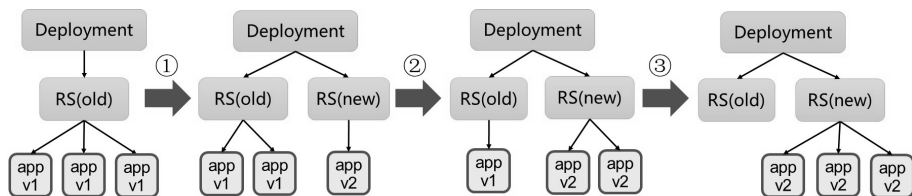


图 2.9 Pod 的滚动升级

下面列出 Deployment nginx-deployment 的详细事件信息：

```
$ kubectl describe deployments/nginx-deployment
Name:          nginx-deployment
Namespace:     default
.....
Replicas:      3 updated | 3 total | 3 available | 0 unavailable
StrategyType:   RollingUpdate
MinReadySeconds: 0
RollingUpdateStrategy: 1 max unavailable, 1 max surge
Conditions:
  Type           Status Reason
  ----           -
  Available      True    MinimumReplicasAvailable
OldReplicaSets:  <none>
NewReplicaSet:   nginx-deployment-3599678771 (3/3 replicas created)
Events:
  FirstSeen    LastSeen    Count   From              SubObjectPath Type
Reason        Message
-----
55m          55m        1   {deployment-controller }      Normal
ScalingReplicaSet Scaled up replica set nginx-deployment-4087004473 to 3
4m           4m         1   {deployment-controller }      Normal
ScalingReplicaSet Scaled up replica set nginx-deployment-3599678771 to 1
4m           4m         1   {deployment-controller }      Normal
ScalingReplicaSet Scaled down replica set nginx-deployment-4087004473 to 2
4m           4m         1   {deployment-controller }      Normal
ScalingReplicaSet Scaled up replica set nginx-deployment-3599678771 to 2
4m           4m         1   {deployment-controller }      Normal
ScalingReplicaSet Scaled down replica set nginx-deployment-4087004473 to 1
4m           4m         1   {deployment-controller }      Normal
ScalingReplicaSet Scaled up replica set nginx-deployment-3599678771 to 3
4m           4m         1   {deployment-controller }      Normal
ScalingReplicaSet Scaled down replica set nginx-deployment-4087004473 to 0
```

运行 `kubectl get rs` 命令，查看两个 ReplicaSet 的最终状态：

```
$ kubectl get rs
```

NAME	DESIRED	CURRENT	READY	AGE
nginx-deployment-3599678771	3	3	3	1m
nginx-deployment-4087004473	0	0	0	52m

在整个升级的过程中，系统会保证至少有两个 Pod 可用，并且最多同时运行 4 个 Pod，这是 Deployment 通过复杂的算法完成的。Deployment 需要确保在整个更新过程中只有一定数量的 Pod 可能处于不可用状态，在默认情况下，Deployment 确保可用的 Pod 总数至少为所需的副本的数量（DESIRED）减 1，也就是最多 1 个不可用（maxUnavailable=1）。Deployment 还需要确保在整个更新过程中 Pod 的总数量不会超过所需的副本数太多，在默认情况下，Deployment 确保 Pod 的总数最多比所需的 Pod 数多 1 个，也就是最多 1 个浪涌值（maxSurge=1）。Kubernetes 从 v1.6 版本开始，maxUnavailable 和 maxSurge 的默认值将从 1、1 更新为所需副本数的 25%、25%。

这样，在升级过程中，Deployment 就能够保证服务不中断，并且副本数量始终维持为用户指定的数量（DESIRED）。

对更新策略的说明如下。

在 Deployment 的定义中，可以通过 spec.strategy 指定 Pod 更新的策略，目前支持两种策略：RollingUpdate（滚动更新）和 Recreate（重建），默认值为 RollingUpdate。在前面的例子中使用的就是 RollingUpdate 策略。

- ◎ Recreate（重建）：设置 spec.strategy.type=Recreate，表示 Deployment 在更新 Pod 时，会先杀掉所有正在运行的 Pod，然后创建新的 Pod。
- ◎ RollingUpdate（滚动更新）：设置 spec.strategy.type=RollingUpdate，表示 Deployment 会以滚动更新的方式来逐个更新 Pod。同时，可以通过设置 spec.strategy.rollingUpdate 下的两个参数（maxUnavailable 和 maxSurge）来控制滚动更新的过程。

RollingUpdate（滚动更新）时两个主要参数的说明如下。

- ◎ spec.strategy.rollingUpdate.maxUnavailable：用于指定 Deployment 在更新过程中不可用状态 Pod 数量的上限。该 maxUnavailable 的数值可以是绝对值（例如 5）或 Pod 期望的副本数的百分比（例如 10%）。如果设置为百分比，那么系统会先以向下取整的方式计算出绝对值（整数）。而当另一个参数 maxSurge 设置为 0 时，maxUnavailable 则必须设置为绝对数值大于 0（从 Kubernetes v1.6 开始，maxUnavailable 的默认值从 1 改为 25%）。举例来说，当 maxUnavailable 设置为 30% 时，旧的 ReplicaSet 可以在滚动更新开始时立即将副本数缩小到所需副本总数的 70%。一旦新的 Pod 创建并准备好，则旧的 ReplicaSet 会进一步缩容，新的 ReplicaSet 又继续扩容，整个过程中系统在任意时刻都可以确保可用状态的 Pod 总数至少占 Pod 期望副本总数的 70%。

- ◎ `spec.strategy.rollingUpdate.maxSurge`: 用于指定 Deployment 更新 Pod 过程中 Pod 总数超过 Pod 期望副本数部分的最大值。该 `maxSurge` 的数值可以是绝对值（例如 5）或 Pod 期望副本数的百分比（例如 10%）。如果设置为百分比，那么系统会先按照向上取整的方式计算出绝对数值（整数）。从 Kubernetes v1.6 开始，`maxSurge` 的默认值从 1 改为 25%。举例来说，当 `maxSurge` 值设置为 30% 时，新的 `ReplicaSet` 可以在滚动更新开始时立即进行副本数扩容，只需要保证新旧 `ReplicaSet` 的 Pod 副本数之和不超过期望副本数的 130% 即可。一旦旧的 Pod 被杀掉，新的 `ReplicaSet` 会进一步扩容。整个过程中系统在任意时刻都能确保新旧 `ReplicaSet` 的 Pod 副本总数之和不超过所需副本数的 130%。

这里需要注意多重更新（Rollover）的情况。如果 Deployment 的上一次更新正在进行，此时用户再次发起 Deployment 的更新操作，那么 Deployment 会为每一次更新都创建一个 `ReplicaSet`，而每次新的 `ReplicaSet` 创建成功后，会逐个增加 Pod 副本数，同时将之前正在扩容的 `ReplicaSet` 停止扩容（更新），并将其加入旧版本 `ReplicaSet` 列表中，然后开始缩容至 0 的操作。

例如，假设我们创建一个 Deployment，这个 Deployment 开始创建 5 个 `Nginx:1.7.9` 的 Pod 副本，在这个创建 Pod 动作尚未完成时，我们又将 Deployment 进行更新，在副本数不变的情况下将 Pod 模板中的镜像修改为 `Nginx:1.9.1`，假设此时 Deployment 已经创建了 3 个 `Nginx:1.7.9` 的 Pod 副本，则 Deployment 会立即杀掉已创建的 3 个 `Nginx:1.7.9` Pod，并开始创建 `Nginx:1.9.1` Pod。Deployment 不会在等待 `Nginx:1.7.9` 的 Pod 创建到 5 个之后再进行更新操作。

还需要注意更新 Deployment 的标签选择器（Label selector）的情况。通常来说，不鼓励更新 Deployment 的标签选择器，因为这样会导致 Deployment 选择的 Pod 列表发生变化，也可能与其他控制器产生冲突。如果一定要更新标签选择器，那么请务必谨慎，确保不会出现其他问题。关于 Deployment 标签选择器的更新的注意事项如下。

（1）添加选择器标签时，必须同步修改 Deployment 配置的 Pod 的标签，为 Pod 添加新的标签，否则 Deployment 的更新会报验证错误而失败：

```
deployments "nginx-deployment" was not valid:
* spec.template.metadata.labels: Invalid value: {"app":"nginx"}: `selector` does not match template `labels`
```

添加标签选择器是无法向后兼容的，这意味着新的标签选择器不会匹配和使用旧选择器创建的 `ReplicaSets` 和 Pod，因此添加选择器将会导致所有旧版本的 `ReplicaSets` 和由旧 `ReplicaSets` 创建的 Pod 处于孤立状态（不会被系统自动删除，也不受新的 `ReplicaSet` 控制）。

为标签选择器和 Pod 模板添加新的标签（使用 `kubect1 edit deployment` 命令）后，效果如下：

```
$ kubect1 get rs
```

NAME	DESIRED	CURRENT	READY	AGE
nginx-deployment-3661742516	3	3	3	2s
nginx-deployment-3599678771	3	3	3	1m
nginx-deployment-4087004473	0	0	0	52m

可以看到新的 **ReplicaSet**（nginx-deployment-3661742516）创建的 3 个新 Pod：

```
$ kubectl get pods
```

NAME	READY	STATUS	RESTARTS	AGE
nginx-deployment-3599678771-01h26	1/1	Running	0	2m
nginx-deployment-3599678771-57thr	1/1	Running	0	2m
nginx-deployment-3599678771-s8p21	1/1	Running	0	2m
nginx-deployment-3661742516-46djm	1/1	Running	0	52s
nginx-deployment-3661742516-kws84	1/1	Running	0	52s
nginx-deployment-3661742516-wq30s	1/1	Running	0	52s

（2）更新标签选择器，即更改选择器中标签的键或者值，也会产生与添加选择器标签类似的效果。

（3）删除标签选择器，即从 **Deployment** 的标签选择器中删除一个或者多个标签，该 **Deployment** 的 **ReplicaSet** 和 **Pod** 不会受到任何影响。但需要注意的是，被删除的标签仍会存在于现有的 **Pod** 和 **ReplicaSets** 上。

## 2. Deployment 的回滚

有时（例如新的 **Deployment** 不稳定时）我们可能需要将 **Deployment** 回滚到旧版本。默认情况下，所有 **Deployment** 的发布历史记录都保留在系统中，以便于我们随时进行回滚（可以配置历史记录数量）。

假设我们在更新 **Deployment** 镜像时，将容器镜像名误设置成 **Nginx:1.91**（一个不存在的镜像）：

```
$ kubectl set image deployment/nginx-deployment nginx=nginx:1.91
deployment "nginx-deployment" image updated
```

则这时 **Deployment** 的部署过程会卡住：

```
$ kubectl rollout status deployments nginx-deployment
Waiting for rollout to finish: 1 out of 3 new replicas have been updated...
```

由于执行过程卡住，所以需要执行 **Ctrl-C** 命令来终止这个查看命令。

查看 **ReplicaSet**，可以看到新建的 **ReplicaSet**（nginx-deployment-3660254150）：

```
$ kubectl get rs
```

NAME	DESIRED	CURRENT	READY	AGE
nginx-deployment-3646295028	3	3	3	53s
<b>nginx-deployment-3660254150</b>	<b>1</b>	<b>1</b>	<b>0</b>	<b>40s</b>
nginx-deployment-4234284026	0	0	0	1m



再查看创建的 Pod，会发现新的 ReplicaSet 创建的 1 个 Pod 卡在镜像拉取过程中。

```
$ kubectl get pods
```

NAME	READY	STATUS	RESTARTS	AGE
nginx-deployment-3646295028-d5r6r	1/1	Running	0	1m
nginx-deployment-3646295028-jw22d	1/1	Running	0	59s
nginx-deployment-3646295028-tw6x7	1/1	Running	0	1m
<b>nginx-deployment-3660254150-9kj51</b>	<b>0/1</b>	<b>ImagePullBackOff</b>	<b>0</b>	<b>49s</b>

为了解决上面这个问题，我们需要回滚到之前稳定版本的 Deployment。

首先，用 `kubectl rollout history` 命令检查这个 Deployment 部署的历史记录：

```
$ kubectl rollout history deployment/nginx-deployment
deployments "nginx-deployment"
```

REVISION	CHANGE-CAUSE
1	kubectl create --filename=nginx-deployment.yaml --record=true
2	kubectl set image deployment/nginx-deployment nginx=nginx:1.9.1
3	kubectl set image deployment/nginx-deployment nginx=nginx:1.9l

注意，在创建 Deployment 时使用 `--record` 参数，就可以在 CHANGE-CAUSE 列看到每个版本使用的命令了。另外，Deployment 的更新操作是在 Deployment 进行部署（Rollout）时被触发的，这意味着当且仅当 Deployment 的 Pod 模板（即 `spec.template`）被更改时才会创建新的修订版本，例如更新模板标签或容器镜像。其他更新操作（如扩展副本数）将不会触发 Deployment 的更新操作，这也意味着我们将 Deployment 回滚到之前的版本时，只有 Deployment 的 Pod 模板部分会被修改。

如果需要查看特定版本的详细信息，则可以加上 `--revision=<N>` 参数：

```
$ kubectl rollout history deployment/nginx-deployment --revision=3
deployments "nginx-deployment" with revision #3
Pod Template:
```

Labels:	app=nginx
	pod-template-hash=3660254150
Annotations:	kubernetes.io/change-cause=kubectl set image deployment/nginx-deployment nginx=nginx:1.9l
Containers:	
nginx:	
Image:	nginx:1.9l
Port:	80/TCP
Environment:	<none>
Mounts:	<none>
Volumes:	<none>

现在我们决定撤销本次发布并回滚到上一个部署版本：

```
$ kubectl rollout undo deployment/nginx-deployment
deployment "nginx-deployment" rolled back
```

当然，也可以使用--to-revision 参数指定回滚到的部署版本号：

```
$ kubectl rollout undo deployment/nginx-deployment --to-revision=2
deployment "nginx-deployment" rolled back
```

这样，该 Deployment 就回滚到之前的稳定版本了，可以从 Deployment 的事件信息中查看到回滚到版本 2 的操作过程：

```
$ kubectl describe deployment/nginx-deployment
Name:                nginx-deployment
.....
OldReplicaSets: <none>
NewReplicaSet:  nginx-deployment-3646295028 (3/3 replicas created)
Events:
  Type      FirstSeen    LastSeen    Count   From              SubObjectPath
  ---      -
  Normal    4m           4m          1      deployment-controller   Scaled up replica set nginx-deployment-4234284026 to 3
  Normal    4m           4m          1      deployment-controller   Scaled up replica set nginx-deployment-3646295028 to 1
  Normal    4m           4m          1      deployment-controller   Scaled down replica set nginx-deployment-4234284026 to 2
  Normal    4m           4m          1      deployment-controller   Scaled up replica set nginx-deployment-3646295028 to 2
  Normal    4m           4m          1      deployment-controller   Scaled down replica set nginx-deployment-4234284026 to 1
  Normal    4m           4m          1      deployment-controller   Scaled up replica set nginx-deployment-3646295028 to 3
  Normal    4m           4m          1      deployment-controller   Scaled down replica set nginx-deployment-4234284026 to 0
  Normal    4m           4m          1      deployment-controller   Scaled up replica set nginx-deployment-3660254150 to 1
  Normal    36s          36s          1      deployment-controller   Rolled back deployment "nginx-deployment" to revision 2
  Normal    36s          36s          1      deployment-controller   Scaled down replica set nginx-deployment-3660254150 to 0
```

### 3. 暂停和恢复 Deployment 的部署操作，以完成复杂的修改

对于一次复杂的 Deployment 配置修改，为了避免频繁触发 Deployment 的更新操作，可以先暂停 Deployment 的更新操作，然后进行配置修改，再恢复 Deployment，一次性触发完整的更新操作，就可以避免不必要的 Deployment 更新操作了。

以之前创建的 Nginx 为例：

```
$ kubectl get deployments
NAME                DESIRED   CURRENT   UP-TO-DATE   AVAILABLE   AGE
nginx-deployment    3         3         0             3           32s
```

```
$ kubectl get rs
NAME                                DESIRED   CURRENT   READY   AGE
nginx-deployment-4234284026        3         3         3       7s
```

通过 `kubectl rollout pause` 命令暂停 Deployment 的更新操作：

```
$ kubectl rollout pause deployment/nginx-deployment
deployment "nginx-deployment" paused
```

然后修改 Deployment 的镜像信息：

```
$ kubectl set image deploy/nginx-deployment nginx=nginx:1.9.1
deployment "nginx-deployment" image updated
```

查看 Deployment 的历史记录，发现并没有触发新的 Deployment 部署操作：

```
$ kubectl rollout history deploy/nginx-deployment
deployments "nginx-deployment"
REVISION    CHANGE-CAUSE
1           kubectl create --filename=nginx-deployment.yaml --record=true
```

在暂停 Deployment 部署之后，可以根据需要进行任意次数的配置更新。例如，再次更新容器的资源限制：

```
$ kubectl set resources deployment nginx-deployment -c=nginx
--limits=cpu=200m,memory=512Mi
deployment "nginx-deployment" resource requirements updated
```

最后，恢复这个 Deployment 的部署操作：

```
$ kubectl rollout resume deploy nginx-deployment
deployment "nginx-deployment" resumed
```

可以看到一个新的 ReplicaSet 被创建出来了：

```
$ kubectl get rs
NAME                                DESIRED   CURRENT   READY   AGE
nginx-deployment-3133440882        3         3         3       6s
nginx-deployment-4234284026        0         0         0       49s
```

查看 **Deployment** 的事件信息，可以看到 **Deployment** 完成了更新：

```
# kubectl describe deployment/nginx-deployment
Name:                nginx-deployment
.....
Events:
  FirstSeen    LastSeen    Count   From              SubObjectPath
Type          Reason      Message
-----
1m            1m          1       deployment-controller
Normal        ScalingReplicaSet  Scaled up replica set nginx-deployment-4234284026
to 3
28s          28s          1       deployment-controller
Normal        ScalingReplicaSet  Scaled up replica set nginx-deployment-3133440882
to 1
27s          27s          1       deployment-controller
Normal        ScalingReplicaSet  Scaled down replica set
nginx-deployment-4234284026 to 2
27s          27s          1       deployment-controller
Normal        ScalingReplicaSet  Scaled up replica set nginx-deployment-3133440882
to 2
26s          26s          1       deployment-controller
Normal        ScalingReplicaSet  Scaled down replica set
nginx-deployment-4234284026 to 1
25s          25s          1       deployment-controller
Normal        ScalingReplicaSet  Scaled up replica set nginx-deployment-3133440882
to 3
23s          23s          1       deployment-controller
Normal        ScalingReplicaSet  Scaled down replica set
nginx-deployment-4234284026 to 0
```

注意，在恢复暂停的 **Deployment** 之前，无法回滚该 **Deployment**。

4. 使用 **kubectl rolling-update** 命令完成 RC 的滚动升级

对于 **RC** 的滚动升级，**Kubernetes** 还提供了一个 **kubectl rolling-update** 命令进行实现。该命令创建了一个新的 **RC**，然后自动控制旧的 **RC** 中的 **Pod** 副本的数量逐渐减少到 0，同时新的 **RC** 中的 **Pod** 副本的数量从 0 逐步增加到目标值，来完成 **Pod** 的升级。需要注意的是，系统要求新的 **RC** 需要与旧的 **RC** 在相同的命名空间（**Namespace**）内，即不能把别人的资产偷偷转移到自家名下。

以 **redis-master** 为例，假设当前运行的 **redis-master Pod** 是 1.0 版本，现在需要升级到 2.0 版本。

创建 `redis-master-controller-v2.yaml` 的配置文件如下：

```
apiVersion: v1
kind: ReplicationController
metadata:
  name: redis-master-v2
  labels:
    name: redis-master
    version: v2
spec:
  replicas: 1
  selector:
    name: redis-master
    version: v2
  template:
    metadata:
      labels:
        name: redis-master
        version: v2
    spec:
      containers:
      - name: master
        image: kubeguide/redis-master:2.0
        ports:
        - containerPort: 6379
```

在配置文件中需要注意以下两点。

- ◎ RC 的名字 (name) 不能与旧的 RC 的名字相同。
- ◎ 在 selector 中应至少有一个 Label 与旧的 RC 的 Label 不同，以标识其为新的 RC。本例中新增了一个名为 version 的 Label，以与旧的 RC 进行区分。

运行 `kubectl rolling-update` 命令完成 Pod 的滚动升级：

```
kubectl rolling-update redis-master -f redis-master-controller-v2.yaml
```

`kubectl` 的执行过程如下：

```
Creating redis-master-v2
At beginning of loop: redis-master replicas: 2, redis-master-v2 replicas: 1
Updating redis-master replicas: 2, redis-master-v2 replicas: 1
At end of loop: redis-master replicas: 2, redis-master-v2 replicas: 1
At beginning of loop: redis-master replicas: 1, redis-master-v2 replicas: 2
Updating redis-master replicas: 1, redis-master-v2 replicas: 2
At end of loop: redis-master replicas: 1, redis-master-v2 replicas: 2
At beginning of loop: redis-master replicas: 0, redis-master-v2 replicas: 3
Updating redis-master replicas: 0, redis-master-v2 replicas: 3
At end of loop: redis-master replicas: 0, redis-master-v2 replicas: 3
Update succeeded. Deleting redis-master
```

```
redis-master-v2
```

等所有新的 Pod 启动完成后，旧的 Pod 也被全部销毁，这样就完成了容器集群的更新工作。

另一种方法是不使用配置文件，直接用 `kubectl rolling-update` 命令，加上 `--image` 参数指定新版镜像名称来完成 Pod 的滚动升级：

```
kubectl rolling-update redis-master --image=redis-master:2.0
```

与使用配置文件的方式不同，执行的结果是旧的 RC 被删除，新的 RC 仍将使用旧的 RC 的名字。

`kubectl` 的执行过程如下：

```
Creating redis-master-ea866a5d2c08588c3375b86fb253db75
At beginning of loop: redis-master replicas: 2, redis-master-ea866a5d2c08588c
3375b86fb253db75 replicas: 1
Updating redis-master replicas: 2, redis-master-ea866a5d2c08588c3375b86fb253db
75 replicas: 1
At end of loop: redis-master replicas: 2, redis-master-ea866a5d2c08588c3375b86fb
253db75 replicas: 1
At beginning of loop: redis-master replicas: 1, redis-master-ea866a5d2c08588c
3375b86fb253db75 replicas: 2
Updating redis-master replicas: 1, redis-master-ea866a5d2c08588c3375b86fb
253db75 replicas: 2
At end of loop: redis-master replicas: 1, redis-master-ea866a5d2c08588c3375b86fb
253db75 replicas: 2
At beginning of loop: redis-master replicas: 0, redis-master-ea866a5d2c08588c
3375b86fb253db75 replicas: 3
Updating redis-master replicas: 0, redis-master-ea866a5d2c08588c3375b86fb253db
75 replicas: 3
At end of loop: redis-master replicas: 0, redis-master-ea866a5d2c08588c3375b86fb
253db75 replicas: 3
Update succeeded. Deleting old controller: redis-master
Renaming redis-master-ea866a5d2c08588c3375b86fb253db75 to redis-master
redis-master
```

可以看到，`kubectl` 通过新建一个新版本 Pod，停掉一个旧版本 Pod，逐步迭代来完成整个 RC 的更新。

更新完成后，查看 RC：

```
$ kubectl get rc
CONTROLLER    CONTAINER(S)   IMAGE(S)           SELECTOR          REPLICAS
redis-master   master         kubeguide/redis-master:2.0    deployment=
ea866a5d2c08588c3375b86fb253db75,name=redis-master,version=v1    3
```

可以看到，`kubectl` 给 RC 增加了一个 key 为 “deployment” 的 Label（这个 key 的名字可通过 `--deployment-label-key` 参数进行修改），Label 的值是 RC 的内容进行 Hash 计算后的值，相当

于签名，这样就能很方便地比较 RC 里的 Image 名字及其他信息是否发生了变化，其具体作用可以参见第 6 章。

如果在更新过程中发现配置有误，则用户可以中断更新操作，并通过执行 `kubectl rolling-update --rollback` 完成 Pod 版本的回滚：

```
$ kubectl rolling-update redis-master --image=kubeguide/redis-master:2.0 --rollback
Found existing update in progress (redis-master-fefd9752aa5883ca4d53013a7b583967), resuming.
Found desired replicas. Continuing update with existing controller redis-master.
At beginning of loop: redis-master-fefd9752aa5883ca4d53013a7b583967 replicas: 0, redis-master replicas: 3
Updating redis-master-fefd9752aa5883ca4d53013a7b583967 replicas: 0, redis-master replicas: 3
At end of loop: redis-master-fefd9752aa5883ca4d53013a7b583967 replicas: 0, redis-master replicas: 3
Update succeeded. Deleting redis-master-fefd9752aa5883ca4d53013a7b583967 redis-master
```

至此，可以看到 Pod 恢复到更新前的版本了。

可以看出，RC 的滚动升级不具有 Deployment 在应用版本升级过程中的历史记录、新旧版本数量的精细控制等功能，在 Kubernetes 的演进过程中，RC 将逐渐被 RS 和 Deployment 所取代，建议用户优先考虑使用 Deployment 完成 Pod 的部署和升级操作。

## 5. 其他管理对象的更新策略

Kubernetes 从 v1.6 版本开始，对 DaemonSet 和 StatefulSet 的更新策略也开始引入类似于 Deployment 的滚动升级，通过不同的策略，自动完成应用的版本升级。

### 1) DaemonSet 的更新策略

目前 DaemonSet 的升级策略包括两种：OnDelete 和 RollingUpdate。

**OnDelete:** DaemonSet 的默认升级策略，与 v1.5 及以前版本的 Kubernetes 保持一致。当使用 OnDelete 作为升级策略时，在创建好新的 DaemonSet 配置之后，新的 Pod 并不会被自动创建，直到用户手动删除旧版本的 Pod，才触发新建操作。

**RollingUpdate:** 从 Kubernetes v1.6 版本开始引入。当使用 RollingUpdate 作为升级策略对 DaemonSet 进行更新时，旧版本的 Pod 将被自动杀掉，然后自动创建新版本的 DaemonSet Pod。整个过程与普通 Deployment 的滚动升级一样是可控的。不过有两点不同于普通 Pod 的滚动升级：一是目前 Kubernetes 还不支持查看和管理 DaemonSet 的更新历史记录；二是 DaemonSet 的回滚 (rollback) 并不能如同 Deployment 一样直接通过 `kubectl rollback` 命令来实现，而是必须通过再次提交旧版本配置的方式实现。

## 2) StatefulSet 的更新策略

Kubernetes 从 v1.6 版本开始, 针对 StatefulSet 的更新策略正逐渐向 Deployment 和 DaemonSet 的更新策略看齐, 也将实现 RollingUpdate、Partitioned 和 OnDelete 几种策略, 目标是保证 StatefulSet 中各 Pod 有序地、逐个地更新, 并且能够保留更新历史, 也能回滚到某个历史版本。

### 2.3.12 Pod 的扩容和缩容

在实际生产系统中, 我们经常会遇到某个服务需要扩容的场景, 也可能会遇到由于资源紧张或者工作负载降低而需要减少服务实例数量的场景。此时我们可以利用 Deployment/RC 的 Scale 机制来完成这些工作。

Kubernetes 对 Pod 的扩容和缩容操作提供了手动和自动两种模式, 手动模式通过执行 `kubectl scale` 命令对一个 Deployment/RC 进行 Pod 副本数量的设置, 即可一键完成。自动模式则需要用户根据某个性能指标或者自定义业务指标, 并指定 Pod 副本数量的范围, 系统将自动在这个范围内根据性能指标的变化进行调整。

#### 1. 手动扩容和缩容模式

以 Deployment nginx 为例:

```
nginx-deployment.yaml
apiVersion: apps/v1beta1
kind: Deployment
metadata:
  name: nginx-deployment
spec:
  replicas: 3
  template:
    metadata:
      labels:
        app: nginx
    spec:
      containers:
      - name: nginx
        image: nginx:1.7.9
        ports:
        - containerPort: 80
```

已运行的 Pod 副本数量为 3 个:

```
$ kubectl get pods
```

NAME	READY	STATUS	RESTARTS	AGE
nginx-deployment-3973253433-scz37	1/1	Running	0	5s



```

nginx-deployment-3973253433-x8fsq 1/1      Running 0      5s
nginx-deployment-3973253433-x9z8z 1/1      Running 0      5s

```

通过 `kubectl scale` 命令可以将 Pod 副本数量从初始的 3 更新为 5:

```

$ kubectl scale deployment nginx-deployment --replicas 5
deployment "nginx-deployment" scaled
$ kubectl get pods

```

NAME	READY	STATUS	RESTARTS	AGE
nginx-deployment-3973253433-3gt27	1/1	Running	0	4s
nginx-deployment-3973253433-7jls2	1/1	Running	0	4s
nginx-deployment-3973253433-scz37	1/1	Running	0	4m
nginx-deployment-3973253433-x8fsq	1/1	Running	0	4m
nginx-deployment-3973253433-x9z8z	1/1	Running	0	4m

将 `--replicas` 设置为比当前 Pod 副本数量更小的数字，系统将会“杀掉”一些运行中的 Pod，以实现应用集群缩容：

```

$ kubectl scale deployment nginx-deployment --replicas=1
deployment "nginx-deployment" scaled

$ kubectl get pods

```

NAME	READY	STATUS	RESTARTS	AGE
nginx-deployment-3973253433-x9z8z	1/1	Running	0	6m

## 2. 自动扩容和缩容模式

从 Kubernetes v1.1 版本开始，新增了名为 Horizontal Pod Autoscaler (HPA) 的控制器，用于实现基于 CPU 使用率进行自动 Pod 扩容和缩容的功能。HPA 控制器基于 Master 的 kube-controller-manager 服务启动参数 `--horizontal-pod-autoscaler-sync-period` 定义的时长（默认值为 30s），周期性地监测目标 Pod 的 CPU 使用率，并在满足条件时对 ReplicationController 或 Deployment 中的 Pod 副本数量进行调整，以符合用户定义的平均 Pod CPU 使用率。Pod CPU 使用率来源于 Heapster 组件，所以需要预先安装好 Heapster。

创建 HPA 时可以使用 `kubectl autoscale` 命令进行快速创建或者使用 yaml 配置文件进行创建。在创建 HPA 之前，需要已经存在一个 Deployment/RC 对象，并且该 Deployment/RC 中的 Pod 必须定义 `resources.requests.cpu` 的资源请求值，如果不设置该值，则 Heapster 将无法采集到该 Pod 的 CPU 使用情况，会导致 HPA 无法正常工作。

下面通过为一个 RC 设置 HPA，然后使用一个客户端对其进行压力测试，对 HPA 的用法进行示例。

以 php-apache 的 RC 为例，设置 cpu request 为 200m，未设置 limit 上限的值：

```

php-apache-deployment.yaml
apiVersion: apps/v1beta1

```

```
kind: Deployment
metadata:
  name: php-apache
spec:
  replicas: 1
  template:
    metadata:
      name: php-apache
      labels:
        app: php-apache
    spec:
      containers:
        - name: php-apache
          image: gcr.io/google_containers/hpa-example
          resources:
            requests:
              cpu: 200m
      ports:
        - containerPort: 80
```

```
# kubectl create -f php-apache-deployment.yaml
deployment "php-apache" created
```

再创建一个 php-apache 的 Service，供客户端访问：

#### **php-apache-svc.yaml**

```
apiVersion: v1
kind: Service
metadata:
  name: php-apache
spec:
  ports:
    - port: 80
  selector:
    app: php-apache
```

```
# kubectl create -f php-apache-svc.yaml
service "php-apache" created
```

接下来为 Deployment “php-apache” 创建一个 HPA 控制器，在 1 和 10 之间调整 Pod 的副本数量，以使得平均 Pod CPU 使用率维持在 50%。

使用 kubectl autoscale 命令进行创建：

```
# kubectl autoscale deployment php-apache --min=1 --max=10 --cpu-percent=50
deployment "php-apache" autoscaled
```

或者通过 yaml 配置文件来创建 HPA，需要在 scaleTargetRef 字段指定需要管理的 Deployment/RC 的名字，然后设置 minReplicas、maxReplicas 和 targetCPUUtilizationPercentage

参数:

```
hpa-php-apache.yaml
apiVersion: autoscaling/v1
kind: HorizontalPodAutoscaler
metadata:
  name: php-apache
spec:
  scaleTargetRef:
    apiVersion: apps/v1beta1
    kind: Deployment
    name: php-apache
  minReplicas: 1
  maxReplicas: 10
  targetCPUUtilizationPercentage: 50

# kubectl create -f hpa-php-apache.yaml
horizontalpodautoscaler "php-apache" created
```

查看已创建的 HPA:

```
# kubectl get hpa
NAME          REFERENCE                TARGETS   MINPODS   MAXPODS   REPLICAS   AGE
php-apache    Deployment/php-apache    1% / 50%   1         10        0          3m
```

然后, 创建一个 busybox Pod, 用于对 php-apache 服务发起压力测试的请求:

```
busybox-pod.yaml
apiVersion: v1
kind: Pod
metadata:
  name: busybox
spec:
  containers:
  - name: busybox
    image: busybox
    command: [ "sleep", "3600" ]

# kubectl create -f busybox-pod.yaml
pod "busybox" created
```

登录 busybox 容器, 执行一个无限循环的 wget 命令来访问 php-apache 服务:

```
# while true; do wget -q -O- http://php-apache > /dev/null; done
```

注意这里 wget 的目的 URL 地址是 Service 的名称 “php-apache”, 这要求 DNS 服务正常工作, 也可以使用 Service 的虚拟 ClusterIP 地址对其进行访问, 例如 http://169.169.122.145:

```
# kubectl exec -ti busybox -- sh
/ # while true; do wget -q -O- http://php-apache > /dev/null; done
```

等待一段时间后，观察 HPA 控制器搜集到的 Pod CPU 使用率：

```
# kubectl get hpa
NAME           REFERENCE                TARGETS  MINPODS  MAXPODS  REPLICAS  AGE
php-apache     Deployment/php-apache    3068%/50%  1        10       0         3m
```

再过一会儿，查看 RC php-apache 副本数量的变化：

```
# kubectl get deployment php-apache
NAME           DESIRED  CURRENT  UP-TO-DATE  AVAILABLE  AGE
php-apache     10       10       10          10         5m
```

可以看到 HPA 已经根据 Pod 的 CPU 使用率的提高对 RC 进行了自动扩容，Pod 的副本数量变成了 10 个。这个过程如图 2.10 所示。

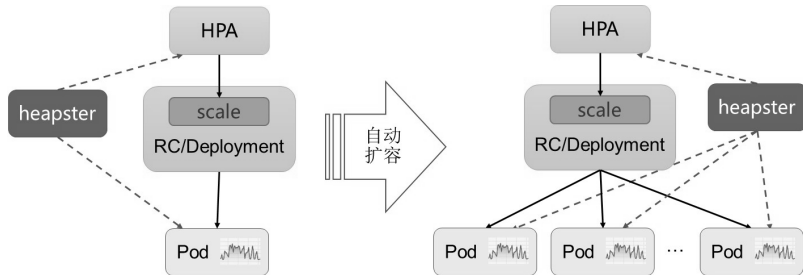


图 2.10 HPA 自动扩容的过程

最后，我们停止压力测试，在 busybox 的控制台输入 Ctrl+C，停止无限循环操作。

等待一段时间，观察 HPA 的变化：

```
# kubectl get hpa
NAME           REFERENCE                TARGETS  MINPODS  MAXPODS  REPLICAS  AGE
php-apache     Deployment/php-apache    3%/50%  1        10       0         10m
```

再次查看 Deployment 的副本数量：

```
# kubectl get deployment php-apache
NAME           DESIRED  CURRENT  UP-TO-DATE  AVAILABLE  AGE
php-apache     1        1        1           1          5m
```

可以看到 HPA 根据 Pod CPU 使用率的降低对副本数量进行了缩容操作，Pod 副本数量变成了 1 个。

当前 HPA 还只支持将 CPU 使用率作为 Pod 副本扩容和缩容的触发条件，在将来的版本中，将会支持应用自定义指标（例如每秒请求数量、请求平均响应时间或其他业务指标）作为触发条件。

### 2.3.13 使用 StatefulSet 搭建 MongoDB 集群

本节以 MongoDB 为例，使用 StatefulSet 完成 MongoDB 集群的创建，为每个 MongoDB 实例在共享存储中（这里采用 GlusterFS）申请一片存储空间，以实现一个无单点故障、高可用、可动态扩展的 MongoDB 集群。部署架构如图 2.11 所示。

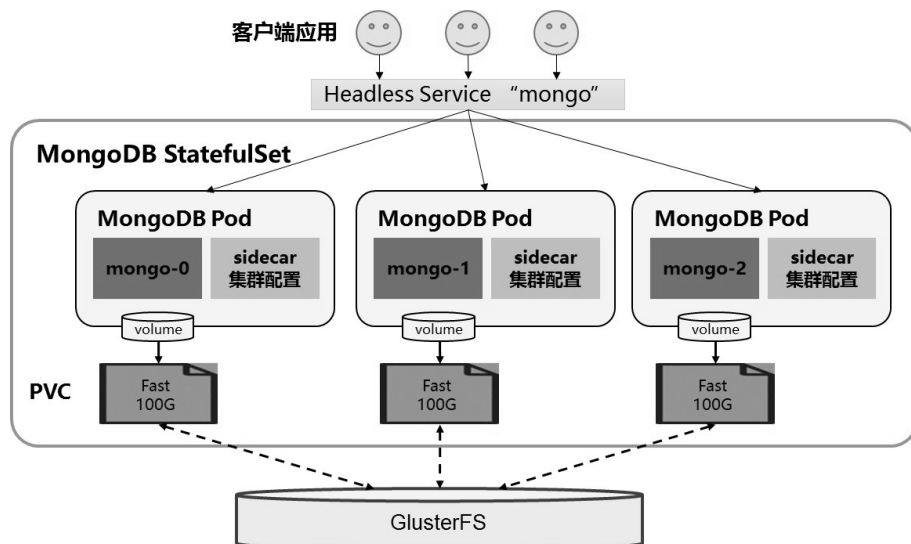


图 2.11 StatefulSet 部署 MongoDB 集群的架构

#### 1. 前提条件

在创建 StatefulSet 之前，需要确保在 Kubernetes 集群中管理员已经创建好共享存储，并能够与 StorageClass 对接，以实现动态存储供应的模式。本节的示例将使用 GlusterFS 作为共享存储（GlusterFS 的部署方法参见 3.8.6 节）。

#### 2. 创建 StatefulSet

为了完成 MongoDB 集群的搭建，需要创建如下三个资源对象。

- ⊙ 一个 StorageClass，用于 StatefulSet 自动为各个应用 Pod 申请 PVC。
- ⊙ 一个 Headless Service，用于维护 MongoDB 集群的状态。
- ⊙ 一个 StatefulSet。

首先，创建一个 StorageClass 对象。

storageclass-fast.yaml 文件的内容如下：

```
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: fast
provisioner: kubernetes.io/glusterfs
parameters:
  resturl: "http://<heketi-rest-url>"
```

执行 `kubecttl create` 命令创建该 StorageClass：

```
# kubecttl create -f storageclass-fast.yaml
storageclass "fast" created
```

接下来，创建对应的 Headless Service。

mongo-sidecar 作为 MongoDB 集群的管理者，将使用此 Headless Service 来维护各个 MongoDB 实例之间的集群关系，以及集群规模变化时的自动更新。

mongo-headless-service.yaml 文件的内容如下：

```
apiVersion: v1
kind: Service
metadata:
  name: mongo
  labels:
    name: mongo
spec:
  ports:
    - port: 27017
      targetPort: 27017
  clusterIP: None
  selector:
    role: mongo
```

使用 `kubecttl create` 命令创建该 StorageClass：

```
# kubecttl create -f mongo-headless-service.yaml
service "mongo" created
```

最后，创建 MongoDB StatefulSet。

statefulset-mongo.yaml 文件的内容如下：

```
apiVersion: apps/v1beta1
kind: StatefulSet
metadata:
  name: mongo
spec:
  serviceName: "mongo"
```

```

replicas: 3
template:
  metadata:
    labels:
      role: mongo
      environment: test
  spec:
    terminationGracePeriodSeconds: 10
    containers:
      - name: mongo
        image: mongo
        command:
          - mongod
          - "--replSet"
          - rs0
          - "--smallfiles"
          - "--noprealloc"
        ports:
          - containerPort: 27017
        volumeMounts:
          - name: mongo-persistent-storage
            mountPath: /data/db
      - name: mongo-sidecar
        image: cvallance/mongo-k8s-sidecar
        env:
          - name: MONGO_SIDECAR_POD_LABELS
            value: "role=mongo,environment=test"
          - name: KUBERNETES_MONGO_SERVICE_NAME
            value: "mongo"
    volumeClaimTemplates:
      - metadata:
          name: mongo-persistent-storage
          annotations:
            volume.beta.kubernetes.io/storage-class: "fast"
        spec:
          accessModes: [ "ReadWriteOnce" ]
          resources:
            requests:
              storage: 100Gi

```

其中的主要配置说明如下。

(1) 在该 `StatefulSet` 的定义中包括两个容器：`mongo` 和 `mongo-sidecar`。`mongo` 是主服务程序，`mongo-sidecar` 是将多个 `mongo` 实例进行集群设置的工具。`mongo-sidecar` 中的环境变量如下。

◎ `MONGO_SIDECAR_POD_LABELS`: 设置为 `mongo` 容器的标签，用于 `sidecar` 查询它

所要管理的 MongoDB 集群实例。

- ◎ KUBERNETES\_MONGO\_SERVICE\_NAME: 它的值为"mongo", 表示 sidecar 将使用 "mongo"这个服务名来完成 MongoDB 集群的设置。

(2) replicas=3 表示这个 MongoDB 集群由 3 个 mongo 实例组成。

(3) volumeClaimTemplates 是 StatefulSet 最重要的存储设置。在 annotations 段设置 volume.beta.kubernetes.io/storage-class="fast"表示使用名为"fast"的 StorageClass 自动为每个 mongo Pod 实例分配后端存储。resources.requests.storage=100Gi 表示为每个 mongo 实例分配 100Gi 的磁盘空间。

使用 kubectl create 命令创建这个 StatefulSet:

```
# kubectl create -f statefulset-mongo.yaml
statefulset "mongo" created
```

最终可以看到 StatefulSet 依次创建并启动了 3 个 mongo Pod 实例, 它们的名字依次为 mongo-0、mongo-1、mongo-2:

```
# kubectl get pods -l role=mongo
NAME          READY   STATUS    RESTARTS   AGE
mongo-0       2/2     Running   0           4m
mongo-1       2/2     Running   0           3m
mongo-2       2/2     Running   0           2m
```

StatefulSet 会用 volumeClaimTemplates 中的定义为每个 Pod 副本创建一个 PVC 实例, 每个 PVC 的名称由 StatefulSet 定义中 volumeClaimTemplates 的名称和 Pod 副本的名字组合而成, 查看系统中的 pvc, 可以验证这一点:

```
# kubectl get pvc
NAME                                STATUS    VOLUME
CAPACITY  ACCESSMODES  STORAGECLASS  AGE
mongo-persistent-storage-mongo-0  Bound
pvc-7d963fef-42b3-11e7-b4ca-000c291bc5fc  100Gi      RWO      fast  4m
mongo-persistent-storage-mongo-1  Bound
pvc-8953f856-42b3-11e7-b4ca-000c291bc5fc  100Gi      RWO      fast  3m
mongo-persistent-storage-mongo-2  Bound
pvc-a0fdc059-42b3-11e7-b4ca-000c291bc5fc  100Gi      RWO      fast  3m
```

下面是 mongo-0 这个 Pod 中的 volumes 设置, 可以看到系统自动为其挂载了对应的 PVC:

```
# kubectl get pod mongo-0 -o yaml
apiVersion: v1
kind: Pod
metadata:
  name: mongo-0
.....
```



```
volumes:
- name: mongo-persistent-storage
  persistentVolumeClaim:
    claimName: mongo-persistent-storage-mongo-0
.....
```

至此，一个由 3 个实例组成的 MongoDB 集群就创建完成了。其中每个实例都拥有稳定的名称和独立的存储空间。

### 3. 查看 MongoDB 集群状态

登录任意一个 mongo Pod，在 mongo 命令行界面用 `rs.status()` 命令查看 MongoDB 集群的状态，可以看到 mongo 集群已通过 sidecar 完成了创建。集群中包含 3 个节点，每个节点的名称都是 StatefulSet 设置的 DNS 域名格式的网络标识名称：

- ◎ mongo-0.mongo.default.svc.cluster.local
- ◎ mongo-1.mongo.default.svc.cluster.local
- ◎ mongo-2.mongo.default.svc.cluster.local

同时，可以看到 3 个 mongo 实例各自的角色（PRIMARY 或 SECONDARY）也都进行了正确的设置。

```
# kubectl exec -ti mongo-0 -- mongo
MongoDB shell version v3.4.4
connecting to: mongodb://127.0.0.1:27017
MongoDB server version: 3.4.4
Welcome to the MongoDB shell.
.....
rs0:PRIMARY>
rs0:PRIMARY> rs.status()
{
  "set" : "rs0",
  "date" : ISODate("2017-05-27T08:13:07.598Z"),
  "myState" : 2,
  "term" : NumberLong(1),
  "syncingTo" : "mongo-0.mongo.default.svc.cluster.local:27017",
  "heartbeatIntervalMillis" : NumberLong(2000),
  "optimes" : {
    "lastCommittedOpTime" : {
      "ts" : Timestamp(1495872747, 1),
      "t" : NumberLong(1)
    },
    "appliedOpTime" : {
      "ts" : Timestamp(1495872747, 1),
      "t" : NumberLong(1)
    }
  }
}
```

```

    },
    "durableOpTime" : {
      "ts" : Timestamp(1495872747, 1),
      "t" : NumberLong(1)
    }
  },
  "members" : [
    {
      "_id" : 0,
      "name" : "mongo-0.mongo.default.svc.cluster.local:27017",
      "health" : 1,
      "state" : 1,
      "stateStr" : "PRIMARY",
      "uptime" : 260,
      "optime" : {
        "ts" : Timestamp(1495872747, 1),
        "t" : NumberLong(1)
      },
      "optimeDurable" : {
        "ts" : Timestamp(1495872747, 1),
        "t" : NumberLong(1)
      },
      "optimeDate" : ISODate("2017-05-27T08:12:27Z"),
      "optimeDurableDate" : ISODate("2017-05-27T08:12:27Z"),
      "lastHeartbeat" : ISODate("2017-05-27T08:13:05.777Z"),
      "lastHeartbeatRecv" :
ISODate("2017-05-27T08:13:05.776Z"),
      "pingMs" : NumberLong(0),
      "electionTime" : Timestamp(1495872445, 1),
      "electionDate" : ISODate("2017-05-27T08:07:25Z"),
      "configVersion" : 9
    },
    {
      "_id" : 1,
      "name" : "mongo-1.mongo.default.svc.cluster.local:27017",
      "health" : 1,
      "state" : 2,
      "stateStr" : "SECONDARY",
      "uptime" : 291,
      "optime" : {
        "ts" : Timestamp(1495872747, 1),
        "t" : NumberLong(1)
      },
      "optimeDate" : ISODate("2017-05-27T08:12:27Z"),
      "syncingTo" : "mongo-0.mongo.default.svc.cluster.local:
27017",
      "configVersion" : 9,

```

```

        "self" : true
    },
    {
        "_id" : 2,
        "name" : "mongo-2.mongo.default.svc.cluster.local:27017",
        "health" : 1,
        "state" : 2,
        "stateStr" : "SECONDARY",
        "uptime" : 164,
        "optime" : {
            "ts" : Timestamp(1495872747, 1),
            "t" : NumberLong(1)
        },
        "optimeDurable" : {
            "ts" : Timestamp(1495872747, 1),
            "t" : NumberLong(1)
        },
        "optimeDate" : ISODate("2017-05-27T08:12:27Z"),
        "optimeDurableDate" : ISODate("2017-05-27T08:12:27Z"),
        "lastHeartbeat" : ISODate("2017-05-27T08:13:06.369Z"),
        "lastHeartbeatRecv" : ISODate("2017-05-27T08:13:06.
635Z"),
        "pingMs" : NumberLong(0),
        "syncingTo" :
"mongo-0.mongo.default.svc.cluster.local:27017",
        "configVersion" : 9
    }
],
"ok" : 1
}

```

对于需要访问这个 mongo 集群的 Kubernetes 集群内部客户端来说，可以通过 Headless Service “mongo” 获取到后端的所有 Endpoints 列表，并组合为数据库链接串，例如 “mongodb://mongo-0.mongo, mongo-1.mongo, mongo-2.mongo:27017/dbname\_?”。

#### 4. StatefulSet 的常见应用场景

下面对 MongoDB 集群常见的两种场景进行操作，说明 StatefulSet 对有状态应用的自动化管理功能。

##### 1) MongoDB 集群的扩容

假设在系统运行过程中，3 个 mongo 实例不足以满足业务的要求，这时就需要对 mongo 集群进行扩容。仅需要通过对 StatefulSet 进行 scale 操作，就能实现在 mongo 集群中自动添加新的 mongo 节点。

使用 `kubect1 scale` 命令将 `StatefulSet` 设置为 4 个实例：

```
# kubect1 scale --replicas=4 statefulset mongo
statefulset "mongo" scaled
```

等待一会儿，看到第 4 个实例 “mongo-3” 创建成功：

```
# kubect1 get po -l role=mongo
NAME          READY    STATUS    RESTARTS   AGE
mongo-0       2/2     Running   0           1h
mongo-1       2/2     Running   0           2h
mongo-2       2/2     Running   0           2h
mongo-3      2/2     Running   0           1m
```

进入某个实例查看 `mongo` 集群的状态，可以看到第 4 个节点已经加入：

```
# kubect1 exec -ti mongo-0 -- mongo
MongoDB shell version v3.4.4
connecting to: mongoddb://127.0.0.1:27017
MongoDB server version: 3.4.4
Welcome to the MongoDB shell.
.....
rs0:PRIMARY>
rs0:PRIMARY> rs.status()
{
.....
  "members" : [
    {
      "_id" : 0,
      "name" : "mongo-0.mongo.default.svc.cluster.local:27017",
      "health" : 1,
      "state" : 1,
      "stateStr" : "PRIMARY",
.....
    {
      "_id" : 4,
      "name" : "mongo-3.mongo.default.svc.cluster.local:27017",
      "health" : 1,
      "state" : 2,
      "stateStr" : "SECONDARY",
      "uptime" : 102,
      "optime" : {
        "ts" : Timestamp(1495880578, 1),
        "t" : NumberLong(4)
      },
      "optimeDurable" : {
        "ts" : Timestamp(1495880578, 1),
        "t" : NumberLong(4)
      },
    },
  ],
}
```

```

        "optimeDate" : ISODate("2017-05-27T10:22:58Z"),
        "optimeDurableDate" : ISODate("2017-05-27T10:22:58Z"),
        "lastHeartbeat" : ISODate("2017-05-27T10:23:00.049Z"),
        "lastHeartbeatRecv" :
ISODate("2017-05-27T10:23:00.049Z"),
        "pingMs" : NumberLong(0),
        "syncingTo" :
"mongo-1.mongo.default.svc.cluster.local:27017",
        "configVersion" : 100097
    }
},
    "ok" : 1
}

```

同时，系统也为 mongo-3 分配了一个新的 PVC 用于保存数据，此处不再赘述，有兴趣的读者可自行查看系统为 mongo-3 绑定的 volume 设置和后端 GlusterFS 共享存储的资源分配情况。

## 2) 自动故障恢复（MongoDB 集群的高可用）

假设在系统运行过程中，某个 mongo 实例或其所在主机发生故障，则 StatefulSet 将会自动重建该 mongo 实例，并保证其身份（ID）和使用的数据（PVC）不变。

以 mongo-0 实例发生故障为例，StatefulSet 将会自动重建 mongo-0 实例，并为其挂载之前分配的 PVC “mongo-persistent-storage-mongo-0”。服务 “mongo-0” 在重新启动后，原数据库中的数据不会丢失，可继续使用。

```

# kubectl get po -l role=mongo
NAME      READY   STATUS             RESTARTS   AGE
mongo-0   0/2     ContainerCreating   0          2h
mongo-1   2/2     Running             0          2h
mongo-2   2/2     Running             0          3s

# kubectl get pod mongo-0 -o yaml
apiVersion: v1
kind: Pod
metadata:
  name: mongo-0
.....
volumes:
- name: mongo-persistent-storage
  persistentVolumeClaim:
    claimName: mongo-persistent-storage-mongo-0
.....

```

进入某个实例查看 mongo 集群的状态，mongo-0 在发生故障前在集群中的角色为 PRIMARY，在其脱离集群后，mongo 集群会自动选出一个 SECONDARY 节点提升为 PRIMARY 节点（本例中为 mongo-2）。重启后的 mongo-0 则会成为一个新的 SECONDARY 节点：

```
# kubectl exec -ti mongo-0 -- mongo
.....
rs0:PRIMARY> rs.status()
{
  .....
    "members" : [
      {
        "_id" : 1,
        "name" : "mongo-1.mongo.default.svc.cluster.local:27017",
        "health" : 1,
        "state" : 2,
        "stateStr" : "SECONDARY",
        .....
      },
      {
        "_id" : 2,
        "name" : "mongo-2.mongo.default.svc.cluster.local:27017",
        "health" : 1,
        "state" : 1,
        "stateStr" : "PRIMARY",
        "uptime" : 6871,
        .....
      },
      {
        "_id" : 3,
        "name" : "mongo-0.mongo.default.svc.cluster.local:27017",
        "health" : 1,
        "state" : 2,
        "stateStr" : "SECONDARY",
        "uptime" : 6806,
        .....
      }
    ]
  }
}
```

从上面的例子中可以看到，Kubernetes 使用 StatefulSet 来搭建有状态的应用集群（MongoDB、MySQL 等），同部署无状态的应用一样简便。Kubernetes 能够保证 StatefulSet 中各应用实例在创建和运行的过程中，都具有固定的身份标识和独立的后端存储；还支持在运行时对集群规模进行扩容、保障集群的高可用等非常重要的功能。

## 2.4 深入掌握 Service

Service 是 Kubernetes 最核心的概念，通过创建 Service，可以为一组具有相同功能的容器应用提供一个统一的入口地址，并且将请求负载分发到后端的各个容器应用上。本节对 Service 的使用进行详细说明，包括 Service 的负载均衡、外网访问、DNS 服务的搭建、Ingress 7 层路由机制等。

## 2.4.1 Service 定义详解

yaml 格式的 Service 定义文件的完整内容如下：

```
apiVersion: v1           // Required
kind: Service            // Required
metadata:                // Required
  name: string           // Required
  namespace: string      // Required
  labels:
    - name: string
  annotations:
    - name: string
spec:                    // Required
  selector: []           // Required
  type: string           // Required
  clusterIP: string
  sessionAffinity: string
  ports:
    - name: string
      protocol: string
      port: int
      targetPort: int
      nodePort: int
  status:
    loadBalancer:
      ingress:
        ip: string
        hostname: string
```

对各属性的说明如表 2.17 所示。

表 2.17 对 Service 的定义文件模板的各属性的说明

属 性 名 称	取 值 类 型	是 否 必 选	取 值 说 明
version	string	Required	v1
kind	string	Required	Service
metadata	object	Required	元数据
metadata.name	string	Required	Service 名称，需符合 RFC 1035 规范
metadata.namespace	string	Required	命名空间，不指定系统时将使用名为“default”的命名空间
metadata.labels[]	list		自定义标签属性列表
metadata.annotation[]	list		自定义注解属性列表
spec	object	Required	详细描述
spec.selector[]	list	Required	Label Selector 配置，将选择具有指定 Label 标签的 Pod 作为管理范围

续表

属 性 名 称	取 值 类 型	是 否 必 选	取 值 说 明
spec.type	string	Required	Service 的类型，指定 Service 的访问方式，默认值为 ClusterIP。 ClusterIP: 虚拟的服务 IP 地址，该地址用于 Kubernetes 集群内部的 Pod 访问，在 Node 上 kube-proxy 通过设置的 Iptables 规则进行转发。 NodePort: 使用宿主机的端口，使能够访问各 Node 的外部客户端通过 Node 的 IP 地址和端口号就能访问服务。 LoadBalancer: 使用外接负载均衡器完成到服务的负载分发，需要在 spec.status.loadBalancer 字段指定外部负载均衡器的 IP 地址，并同时定义 nodePort 和 clusterIP，用于公有云环境
spec.clusterIP	string		虚拟服务 IP 地址，当 type=ClusterIP 时，如果不指定，则系统进行自动分配，也可以手工指定；当 type=LoadBalancer 时，则需要指定
spec.sessionAffinity	string		是否支持 Session，可选值为 ClientIP，默认值为空。 ClientIP: 表示将同一个客户端（根据客户端的 IP 地址决定）的访问请求都转发到同一个后端 Pod
spec.ports[]	list		Service 需要暴露的端口列表
spec.ports[].name	string		端口名称
spec.ports[].protocol	string		端口协议，支持 TCP 和 UDP，默认值为 TCP
spec.ports[].port	int		服务监听的端口号
spec.ports[].targetPort	int		需要转发到后端 Pod 的端口号
spec.ports[].nodePort	int		当 spec.type=NodePort 时，指定映射到物理机的端口号
Status	object		当 spec.type=LoadBalancer 时，设置外部负载均衡器的地址，用于公有云环境
status.loadBalancer	object		外部负载均衡器
status.loadBalancer.ingress	object		外部负载均衡器
status.loadBalancer.ingress.ip	string		外部负载均衡器的 IP 地址
status.loadBalancer.ingress.hostname	string		外部负载均衡器的主机名

2.4.2 Service 基本用法

一般来说，对外提供服务的应用程序需要通过某种机制来实现，对于容器应用最简便的方式就是通过 TCP/IP 机制及监听 IP 和端口号来实现。例如，我们定义一个提供 Web 服务的 RC，由两个 tomcat 容器副本组成，每个容器通过 containerPort 设置提供服务的端口号为 8080：

```
webapp-rc.yaml
apiVersion: v1
kind: ReplicationController
```



```

metadata:
  name: webapp
spec:
  replicas: 2
  template:
    metadata:
      name: webapp
      labels:
        app: webapp
    spec:
      containers:
      - name: webapp
        image: tomcat
        ports:
          - containerPort: 8080

```

创建该 RC:

```

# kubectl create -f webapp-rc.yaml
replicationcontroller "webapp" created

```

获取 Pod 的 IP 地址:

```

# kubectl get pods -l app=webapp -o yaml | grep podIP
  podIP: 172.17.1.4
  podIP: 172.17.1.3

```

可以直接通过这两个 Pod 的 IP 地址和端口号访问 Tomcat 服务:

```

# curl 172.17.1.3:8080
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="UTF-8" />
    <title>Apache Tomcat/8.0.35</title>
  </head>
  <body>
    <h1>Apache Tomcat/8.0.35</h1>
  </body>
</html>
.....
# curl 172.17.1.4:8080
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="UTF-8" />
    <title>Apache Tomcat/8.0.35</title>
  </head>
  <body>
    <h1>Apache Tomcat/8.0.35</h1>
  </body>
</html>
.....

```

直接通过 Pod 的 IP 地址和端口号可以访问到容器应用内的服务，但是 Pod 的 IP 地址是不可靠的，例如当 Pod 所在的 Node 发生故障时，Pod 将被 Kubernetes 重新调度到另一台 Node，Pod 的 IP 地址将发生变化。更重要的是，如果容器应用本身是分布式的部署方式，通过多个实例共同提供服务，就需要在这些实例的前端设置一个负载均衡器来实现请求的分发。Kubernetes 中的 Service 就是设计出来用于解决这些问题的核心组件。

以前面创建的 `webapp` 应用为例，为了让客户端应用能够访问到两个 Tomcat Pod 实例，需要创建一个 `Service` 来提供服务。Kubernetes 提供了一种快速的方法，即通过 `kubectl expose` 命令来创建 `Service`：

```
# kubectl expose rc webapp
service "webapp" exposed
```

查看新创建的 `Service`，可以看到系统为它分配了一个虚拟的 IP 地址（`ClusterIP`），而 `Service` 所需的端口号则从 Pod 中的 `containerPort` 复制而来：

```
# kubectl get svc
NAME          CLUSTER-IP      EXTERNAL-IP      PORT(S)          AGE
webapp        169.169.235.79  <none>           8080/TCP         3s
```

接下来，我们就可以通过 `Service` 的 IP 地址和 `Service` 的端口号访问该 `Service` 了：

```
# curl 169.169.235.79:8080
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="UTF-8" />
    <title>Apache Tomcat/8.0.35</title>
  </head>
  <body>
    <h1>Apache Tomcat/8.0.35</h1>
  </body>
</html>
.....
```

这里，对 `Service` 地址 `169.169.235.79:8080` 的访问被自动负载分发到了后端两个 Pod 之一：`172.17.1.3:8080` 或 `172.17.1.4:8080`。

除了使用 `kubectl expose` 命令创建 `Service`，我们也可以通过配置文件定义 `Service`，再通过 `kubectl create` 命令进行创建。例如对于前面的 `webapp` 应用，我们可以设置一个 `Service`，代码如下：

```
apiVersion: v1
kind: Service
metadata:
  name: webapp
spec:
  ports:
  - port: 8081
    targetPort: 8080
  selector:
    app: webapp
```

`Service` 定义中的关键字段是 `ports` 和 `selector`。本例中 `ports` 定义部分指定了 `Service` 所需的虚拟端口号为 `8081`，由于与 Pod 容器端口号 `8080` 不一样，所以需要再通过 `targetPort` 来指定后端 Pod 的端口号。`selector` 定义部分设置的是后端 Pod 所拥有的是 `label: app=webapp`。

创建该 `Service` 并查看其 `ClusterIP` 地址：

```
# kubectl create -f webapp-svc.yaml
service "webapp" created
```

```
# kubectl get svc
NAME          CLUSTER-IP      EXTERNAL-IP  PORT(S)    AGE
webapp        169.169.28.190  <none>       8081/TCP    3s
```

通过 Service 的 IP 地址和 Service 的端口号进行访问：

```
# curl 169.169.28.190:8081
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="UTF-8" />
    <title>Apache Tomcat/8.0.35</title>
  </head>
  <body>
    <h1>Apache Tomcat/8.0.35</h1>
  </body>
</html>
.....
```

同样，对 Service 地址 169.169.28.190:8081 的访问被自动负载分发到了后端两个 Pod 之一：172.17.1.3:8080 或 172.17.1.4:8080。目前 Kubernetes 提供了两种负载分发策略：RoundRobin 和 SessionAffinity，具体说明如下。

- ◎ RoundRobin：轮询模式，即轮询将请求转发到后端的各个 Pod 上。
- ◎ SessionAffinity：基于客户端 IP 地址进行会话保持的模式，即第 1 次将某个客户端发起的请求转发到后端的某个 Pod 上，之后从相同的客户端发起的请求都将被转发到后端相同的 Pod 上。

在默认情况下，Kubernetes 采用 RoundRobin 模式对客户端请求进行负载分发，但我们也可以通过设置 `service.spec.sessionAffinity=ClientIP` 来启用 SessionAffinity 策略。这样，同一个客户端 IP 发来的请求就会被转发到后端固定的某个 Pod 上了。

通过 Service 的定义，Kubernetes 实现了一种分布式应用统一入口的定义和负载均衡机制。Service 还可以进行其他类型的设置，例如设置多个端口号、直接设置为集群外部服务，或实现为无头服务（Headless）模式（将在 2.4.3 节介绍）。

## 1. 多端口 Service

有时一个容器应用也可能提供多个端口的服务，那么在 Service 的定义中也可以相应地设置为将多个端口对应到多个应用服务。在下面的例子中，Service 设置了两个端口号，并且为每个端口号进行了命名：

```
apiVersion: v1
kind: Service
metadata:
  name: webapp
```

```
spec:
  ports:
    - port: 8080
      targetPort: 8080
      name: web
    - port: 8005
      targetPort: 8005
      name: management
  selector:
    app: webapp
```

另一个例子是两个端口号使用了不同的 4 层协议——TCP 和 UDP：

```
apiVersion: v1
kind: Service
metadata:
  name: kube-dns
  namespace: kube-system
  labels:
    k8s-app: kube-dns
    kubernetes.io/cluster-service: "true"
    kubernetes.io/name: "KubeDNS"
spec:
  selector:
    k8s-app: kube-dns
  clusterIP: 169.169.0.100
  ports:
    - name: dns
      port: 53
      protocol: UDP
    - name: dns-tcp
      port: 53
      protocol: TCP
```

## 2. 外部服务 Service

在某些环境中，应用系统需要将一个外部数据库作为后端服务进行连接，或将另一个集群或 Namespace 中的服务作为服务的后端，这时可以通过创建一个无 Label Selector 的 Service 来实现：

```
kind: Service
apiVersion: v1
metadata:
  name: my-service
spec:
  ports:
    - protocol: TCP
```

```
port: 80
targetPort: 80
```

通过该定义创建的是一个不带标签选择器的 Service，即无法选择后端的 Pod，系统不会自动创建 Endpoint，因此需要手动创建一个和该 Service 同名的 Endpoint，用于指向实际的后端访问地址。创建 Endpoint 的配置文件内容如下：

```
kind: Endpoints
apiVersion: v1
metadata:
  name: my-service
subsets:
- addresses:
  - IP: 1.2.3.4
  ports:
  - port: 80
```

如图 2.12 所示，访问没有标签选择器的 Service 和带有标签选择器的 Service 一样，请求将会被路由到由用户手动定义的后端 Endpoint 上。

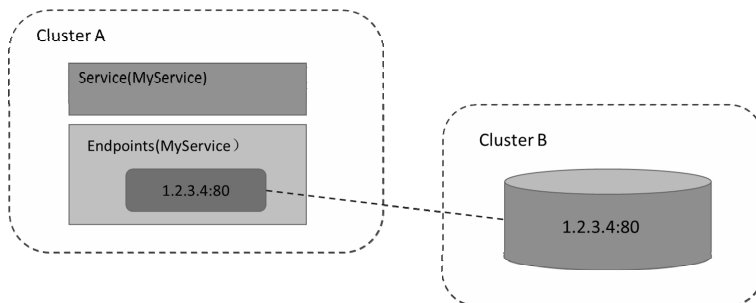


图 2.12 Service 指向外部服务

### 2.4.3 Headless Service

在某些应用场景中，开发人员希望自己控制负载均衡的策略，不使用 Service 提供的默认负载均衡的功能，或者应用程序希望知道属于同组服务的其他实例。Kubernetes 提供了 Headless Service（无头服务）来实现这种功能，即不为 Service 设置 ClusterIP（入口 IP 地址），仅通过 Label Selector 将后端的 Pod 列表返回给调用的客户端。例如：

```
apiVersion: v1
kind: Service
metadata:
  name: nginx
  labels:
    app: nginx
```

```
spec:
  ports:
  - port: 80
  clusterIP: None
  selector:
    app: nginx
```

这样，Service 就不再具有一个特定的 ClusterIP 地址，对其进行访问将获得包含 Label “app=nginx”的全部 Pod 列表，然后客户端程序自行决定如何处理这个 Pod 列表。例如，StatefulSet 就是使用 Headless Service 为客户端返回多个服务地址。

对于“去中心化”类的应用集群，Headless Service 将非常有用。下面以搭建 Cassandra 集群为例，看看如何通过对 Headless Service 的巧妙使用，自动实现应用集群的创建。

Apache Cassandra 是一套开源分布式 NoSQL 数据库系统，主要特点为它不是单个数据库，而是由一组数据库节点共同构成的一个分布式的集群数据库。由于 Cassandra 使用的是“去中心化”模式，所以在集群里的一个节点启动之后，需要一个途径获知集群中新节点的加入。Cassandra 使用了 Seed（种子）来完成在集群中节点之间的相互查找和通信。

通过对 Headless Service 的使用，实现了 Cassandra 各节点之间的相互查找和集群的自动搭建。主要步骤包括：自定义 SeedProvider；通过 Headless Service 自动查找后端 Pod；自动添加新 Cassandra 节点。

## 1. 自定义 SeedProvider

在本例中使用了一个自定义的 SeedProvider 类来完成新节点的查询和添加，类名为 io.k8s.cassandra.KubernetesSeedProvider。

KubernetesSeedProvider.java 类的源代码节选如下：

```
.....
public List<InetAddress> getSeeds() {
    List<InetAddress> list = new ArrayList<InetAddress>();
    String host = "https://kubernetes.default.cluster.local";
    String serviceName = getEnvOrDefault("CASSANDRA_SERVICE", "cassandra");
    String podNamespace = getEnvOrDefault("POD_NAMESPACE", "default");
    String path = String.format("/api/v1/namespaces/%s/endpoints/", podNamespace);
    .....
    public static void main(String[] args) {
        SeedProvider provider = new KubernetesSeedProvider(new HashMap<String,
String>());
        System.out.println(provider.getSeeds());
    }
}
```

完整的源代码可以从 <https://github.com/kubernetes/kubernetes/blob/master/examples/storage/cassandra/java/src/main/java/io/k8s/cassandra/KubernetesSeedProvider.java> 获取。

定制的 `KubernetesSeedProvider` 类将使用 REST API 来访问 Kubernetes Master，然后通过查询 `name=cassandra` 的服务（Headless Service 将返回 Pod 列表）完成对其他“节点”的查找。

## 2. 通过 Service 动态查找 Pod

在 `KubernetesSeedProvider` 类中，通过查询环境变量 `CASSANDRA_SERVICE` 的值来获得服务的名称。这样就要求 Service 需要在 Pod 之前创建出来。如果我们已经创建好 DNS 服务（参见 2.5.5 节的案例介绍），那么也可以直接使用服务的名称而无须使用环境变量。

回顾一下 Service 的概念。Service 通常用作一个负载均衡器，供 Kubernetes 集群中其他应用（Pod）对属于该 Service 的一组 Pod 进行访问。由于 Pod 的创建和销毁都会实时更新 Service 的 Endpoints 数据，所以可以动态地对 Service 的后端 Pod 进行查询了。Cassandra 的“去中心化”设计使得 Cassandra 集群中的一个 Cassandra 实例（节点）只需要查询到其他节点，即可自动组成一个集群，正好可以使用 Service 的这个特性查询到新增的节点。图 2.13 描述了 Cassandra 新节点加入集群的过程。

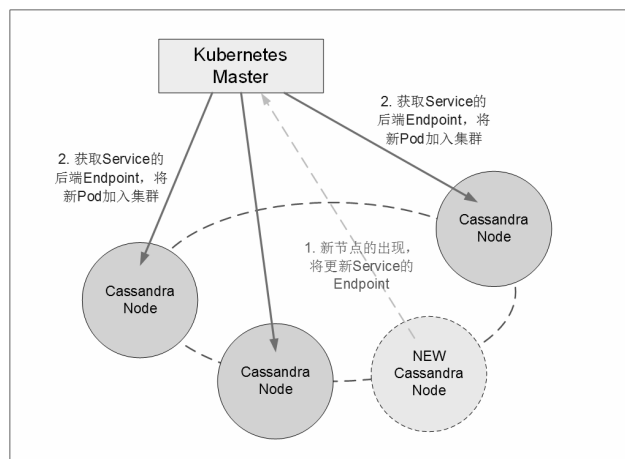


图 2.13 Cassandra 新节点加入集群的过程

在 Kubernetes 系统中，首先需要为 Cassandra 集群定义一个 Service。

### cassandra-service.yaml

```
apiVersion: v1
kind: Service
metadata:
  labels:
```

```
    name: cassandra
  name: cassandra
spec:
  ports:
    - port: 9042
  selector:
    name: cassandra
```

在 Service 的定义中指定 Label Selector 为 name=cassandra。

使用 `kubecttl create` 命令创建这个 Service：

```
$ kubecttl create -f cassandra-service.yaml
service "cassandra" created
```

然后，创建一个 Cassandra Pod：

### **cassandra-rc.yaml**

```
apiVersion: v1
kind: ReplicationController
metadata:
  labels:
    name: cassandra
  name: cassandra
spec:
  replicas: 1
  selector:
    name: cassandra
  template:
    metadata:
      labels:
        name: cassandra
    spec:
      containers:
        - command:
            - /run.sh
          resources:
            limits:
              cpu: 0.5
          env:
            - name: MAX_HEAP_SIZE
              value: 512M
            - name: HEAP_NEWSIZE
              value: 100M
            - name: POD_NAMESPACE
              valueFrom:
                fieldRef:
                  fieldPath: metadata.namespace
```



```

image: gcr.io/google_containers/cassandra:v5
name: cassandra
ports:
  - containerPort: 9042
    name: cql
  - containerPort: 9160
    name: thrift
volumeMounts:
  - mountPath: /cassandra_data
    name: data
volumes:
  - name: data
    emptyDir: {}

```

```

$ kubectl create -f cassandra-rc.yaml
replicationcontroller "cassandra" created

```

现在，一个 Cassandra Pod 运行起来了，但还没有组成 Cassandra 集群。

### 3. Cassandra 集群中新节点的自动添加

现在，我们使用 Kubernetes 提供的 Scale（扩容和缩容）机制对 Cassandra 集群进行扩容：

```

$ kubectl scale rc cassandra --replicas=2
replicationcontroller "cassandra" scaled

```

查看 Pod，可以看到 RC 创建并启动了一个新的 Pod：

```

$ kubectl get pods -l="name=cassandra"

```

NAME	READY	STATUS	RESTARTS	AGE
cassandra	1/1	Running	0	5m
cassandra-g52t3	1/1	Running	0	50s

使用 Cassandra 提供的 `nodetool` 工具对任一 `cassandra` 实例(Pod)进行访问来验证 Cassandra 集群的状态。下面的命令将访问名为 `cassandra` 的 Pod（访问 `cassandra-g52t3` 也能获得相同的结果）：

```

$ kubectl exec -ti cassandra -- nodetool status
Datacenter: datacenter1
=====
Status=Up/Down
|/ State=Normal/Leaving/Joining/Moving
-- Address      Load      Tokens   Owns (effective)  Host ID                               Rack
UN  10.1.20.16    51.58 KB   256      100.0%            1625c65d-b5b6-40f4-a794-6f5a12322d86 rack1
UN  10.1.10.11    51.51 KB   256      100.0%            cdfcbf1a-795c-4412-9d3f-e8fe50bb8deb rack1

```

可以看到 Cassandra 集群中有两个节点处于正常运行状态（Up and Normal，UN）。该结果

中的两个 IP 地址为两个 Cassandra Pod 的 IP 地址。

内部的过程为：每个 Cassandra 节点（Pod）通过 API 访问 Kubernetes Master，查询名为 `cassandra` 的 Service 的 Endpoints（即 Cassandra 节点），若发现有新节点加入，就进行添加操作，最后成功组成了一个 Cassandra 集群。

我们再增加两个 Cassandra 实例：

```
$ kubectl scale rc cassandra --replicas=4
```

用 `nodetool` 工具查看 Cassandra 集群状态：

```
$ kubectl exec -ti cassandra -- nodetool status
Datacenter: datacenter1
=====
Status=Up/Down
|/ State=Normal/Leaving/Joining/Moving
-- Address      Load          Tokens   Owns (effective)  Host ID                                     Rack
UN  10.1.20.16   51.58 KB      256      50.5%             1625c65d-b5b6-40f4-a794-6f5a12322d86 rack1
UN  10.1.10.12   52.03 KB      256      47.0%             8bcc1c3e-44ec-46a7-b981-4090b206f14e rack1
UN  10.1.20.17   68.05 KB      256      50.6%             579b6493-e92a-47f5-91f2-9313198a24c9 rack1
UN  10.1.10.11   51.51 KB      256      51.9%             cdfcbf1a-795c-4412-9d3f-e8fe50bb8deb rack1
```

可以看到 4 个 Cassandra 节点都加入 Cassandra 集群中了。

另外，可以通过查看 Cassandra Pod 的日志来看到新节点加入集群的记录：

```
$ kubectl logs cassandra-g52t3
.....
INFO 18:05:36 Handshaking version with /10.1.20.17
INFO 18:05:36 Node /10.1.20.17 is now part of the cluster
INFO 18:05:36 InetAddress /10.1.20.17 is now UP
INFO 18:05:38 Handshaking version with /10.1.10.12
INFO 18:05:39 Node /10.1.10.12 is now part of the cluster
INFO 18:05:39 InetAddress /10.1.10.12 is now UP
```

本例描述了一种通过 API 查询 Service 来完成动态 Pod 发现的应用场景。对于类似于 Cassandra 的去中心化集群类应用，都可以使用 Headless Service 查询后端 Endpoints 这种巧妙的方法来实现对应用集群（属于同一 Service）中新加入节点的查找。

## 2.4.4 集群外部访问 Pod 或 Service

由于 Pod 和 Service 是 Kubernetes 集群范围内的虚拟概念，所以集群外的客户端系统无法通

过 Pod 的 IP 地址或者 Service 的虚拟 IP 地址和虚拟端口号访问到它们。为了让外部客户端可以访问这些服务，可以将 Pod 或 Service 的端口号映射到宿主机，以使得客户端应用能够通过物理机访问容器应用。

## 1. 将容器应用的端口号映射到物理机

(1) 通过设置容器级别的 hostPort，将容器应用的端口号映射到物理机上：

```
pod-hostport.yaml
apiVersion: v1
kind: Pod
metadata:
  name: webapp
  labels:
    app: webapp
spec:
  containers:
  - name: webapp
    image: tomcat
    ports:
      - containerPort: 8080
        hostPort: 8081
```

通过 `kubectl create` 命令创建这个 Pod：

```
# kubectl create -f pod-hostport.yaml
pod "webapp" created
```

通过物理机的 IP 地址和 8081 端口号访问 Pod 内的容器服务：

```
# curl 192.168.18.3:8081
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="UTF-8" />
    <title>Apache Tomcat/8.0.35</title>
.....
```

(2) 通过设置 Pod 级别的 `hostNetwork=true`，该 Pod 中所有容器的端口号都将被直接映射到物理机上。设置 `hostNetwork=true` 时需要注意，在容器的 `ports` 定义部分如果不指定 `hostPort`，则默认 `hostPort` 等于 `containerPort`，如果指定了 `hostPort`，则 `hostPort` 必须等于 `containerPort` 的值。

```
pod-hostnetwork.yaml
apiVersion: v1
kind: Pod
metadata:
```



See <http://releases.k8s.io/release-1.3/docs/user-guide/services-firewalls.md> for more details.

```
service "webapp" created
```

系统提示信息说明：由于要使用物理机的端口号，所以需要在防火墙上做好相应的配置，以使得外部客户端能够访问到该端口。

通过物理机的 IP 地址和 nodePort 8081 端口号访问服务:

```
# curl 192.168.18.3:8081
<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="UTF-8" />
    <title>Apache Tomcat/8.0.35</title>
.....
```

同样，对该 Service 的访问也将被负载分发到后端的多个 Pod 上。

(2) 通过设置 `LoadBalancer` 映射到云服务商提供的 `LoadBalancer` 地址。这种用法仅用于在公有云服务提供商的云平台上设置 `Service` 的场景。在下面的例子中，`status.loadBalancer.ingress.ip` 设置的 `146.148.47.155` 为云服务商提供的负载均衡器的 IP 地址。对该 `Service` 的访问请求将会通过 `LoadBalancer` 转发到后端 `Pod` 上，负载分发的实现方式则依赖于云服务商提供的 `LoadBalancer` 的实现机制。

```
kind: Service
apiVersion: v1
metadata:
  name: my-service
spec:
  selector:
    app: MyApp
  ports:
    - protocol: TCP
      port: 80
      targetPort: 9376
      nodePort: 30061
  clusterIP: 10.0.171.239
  loadBalancerIP: 78.11.24.19
  type: LoadBalancer
status:
  loadBalancer:
    ingress:
      - ip: 146.148.47.155
```

### 2.4.5 DNS 服务搭建指南

作为服务发现机制的基本功能，在集群内需要能够通过服务名对服务进行访问，这就需要 一个集群范围的 DNS 服务来完成服务名到 ClusterIP 的解析。本节将对如何搭建 DNS 服务进行 详细说明。

Kubernetes 提供的虚拟 DNS 服务名为 skydns，由 4 个组件组成。

- (1) etcd: DNS 存储。
- (2) kube2sky: 将 Kubernetes Master 中的 Service（服务）注册到 etcd。
- (3) skyDNS: 提供 DNS 域名解析服务。
- (4) healthz: 提供对 skydns 服务的健康检查功能。

图 2.14 描述了 Kubernetes DNS 服务的总体架构。

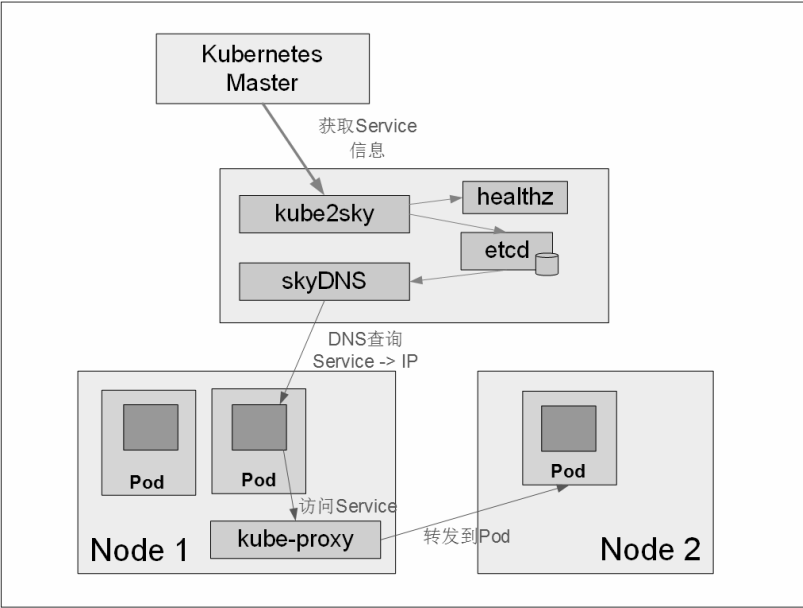


图 2.14 Kubernetes DNS 服务的总体架构

#### 1. skydns 配置文件说明

skydns 服务由一个 RC 和一个 Service 的定义组成，分别由配置文件 skydns-rc.yaml 和 skydns-svc.yaml 定义。

skydns 的 RC 配置文件 skydns-rc.yaml 的内容如下，包含了 4 个容器的定义：

```
skydns-rc.yaml
apiVersion: v1
kind: ReplicationController
metadata:
  name: kube-dns-v11
  namespace: kube-system
  labels:
    k8s-app: kube-dns
    version: v11
    kubernetes.io/cluster-service: "true"
spec:
  replicas: 1
  selector:
    k8s-app: kube-dns
    version: v11
  template:
    metadata:
      labels:
        k8s-app: kube-dns
        version: v11
        kubernetes.io/cluster-service: "true"
    spec:
      containers:
        - name: etcd
          image: gcr.io/google_containers/etcd-amd64:2.2.1
          resources:
            limits:
              cpu: 100m
              memory: 50Mi
            requests:
              cpu: 100m
              memory: 50Mi
          command:
            - /usr/local/bin/etcd
            - -data-dir
            - /tmp/data
            - -listen-client-urls
            - http://127.0.0.1:2379,http://127.0.0.1:4001
            - -advertise-client-urls
            - http://127.0.0.1:2379,http://127.0.0.1:4001
            - -initial-cluster-token
            - skydns-etcd
          volumeMounts:
            - name: etcd-storage
              mountPath: /tmp/data
```

```
- name: kube2sky
image: gcr.io/google_containers/kube2sky-amd64:1.15
resources:
  limits:
    cpu: 100m
    # Kube2sky watches all pods.
    memory: 50Mi
  requests:
    cpu: 100m
    memory: 50Mi
livenessProbe:
  httpGet:
    path: /healthz
    port: 8080
    scheme: HTTP
  initialDelaySeconds: 60
  timeoutSeconds: 5
  successThreshold: 1
  failureThreshold: 5
readinessProbe:
  httpGet:
    path: /readiness
    port: 8081
    scheme: HTTP
  # we poll on pod startup for the Kubernetes master service and
  # only setup the /readiness HTTP server once that's available.
  initialDelaySeconds: 30
  timeoutSeconds: 5
args:
  # command = "/kube2sky"
  - --kube-master-url=http://192.168.18.3:8080
  - --domain=cluster.local
- name: skydns
image: gcr.io/google_containers/skydns:2015-10-13-8c72f8c
resources:
  limits:
    cpu: 100m
    memory: 50Mi
  requests:
    cpu: 100m
    memory: 50Mi
args:
  # command = "/skydns"
  - -machines=http://127.0.0.1:4001
  - -addr=0.0.0.0:53
  - -ns-rotate=false
  - --domain=cluster.local
```



```

ports:
- containerPort: 53
  name: dns
  protocol: UDP
- containerPort: 53
  name: dns-tcp
  protocol: TCP
- name: healthz
  image: gcr.io/google_containers/exechealthz:1.0
  resources:
    # keep request = limit to keep this container in guaranteed class
    limits:
      cpu: 10m
      memory: 20Mi
    requests:
      cpu: 10m
      memory: 20Mi
  args:
  - -cmd=nslookup kubernetes.default.svc.cluster.local 127.0.0.1 >/dev/null
  - -port=8080
  ports:
  - containerPort: 8080
    protocol: TCP
  volumes:
  - name: etcd-storage
    emptyDir: {}
  dnsPolicy: Default # Don't use cluster DNS.

```

需要修改的几个配置参数如下。

(1) kube2sky 容器需要访问 Kubernetes Master，需要配置 Master 所在物理主机的 IP 地址和端口号，本例中设置参数--kube\_master\_url 的值为 http://192.168.18.3:8080。

(2) kube2sky 容器和 skydns 容器的启动参数--domain，设置 Kubernetes 集群中 Service 所属的域名，本例中为“cluster.local”。启动后，kube2sky 会通过 API Server 监控集群中全部 Service 的定义，生成相应的记录并保存到 etcd 中。kube2sky 为每个 Service 生成以下两条记录。

- ◎ <service\_name>.<namespace\_name>.<domain>。
- ◎ <service\_name>.<namespace\_name>.svc.<domain>。

(3) skydns 的启动参数--addr=0.0.0.0:53 表示使用本机 TCP 和 UDP 的 53 端口提供服务。

skydns 的 Service 配置文件 skydns-svc.yaml 的内容如下：

```

skydns-svc.yaml
apiVersion: v1
kind: Service

```

```
metadata:
  name: kube-dns
  namespace: kube-system
  labels:
    k8s-app: kube-dns
    kubernetes.io/cluster-service: "true"
    kubernetes.io/name: "KubeDNS"
spec:
  selector:
    k8s-app: kube-dns
  clusterIP: 169.169.0.100
  ports:
    - name: dns
      port: 53
      protocol: UDP
    - name: dns-tcp
      port: 53
      protocol: TCP
```

注意，skydns 服务使用的 clusterIP 需要我们指定一个固定的 IP 地址，每个 Node 的 kubelet 进程都将使用这个 IP 地址，不能通过 Kubernetes 自动分配。

另外，这个 IP 地址需要在 kube-apiserver 启动参数--service-cluster-ip-range 指定的 IP 地址范围内。

在创建 skydns 容器之前，先修改每个 Node 上 kubelet 的启动参数。

## 2. 修改每台 Node 上的 kubelet 启动参数

修改每台 Node 上 kubelet 的启动参数，加上以下两个参数。

- ◎ --cluster\_dns=169.169.0.100: 为 DNS 服务的 ClusterIP 地址。
- ◎ --cluster\_domain=cluster.local: 为 DNS 服务中设置的域名。

然后重启 kubelet 服务。

## 3. 创建 skydns RC 和 Service

通过 kubectl create 完成 skydns 的 RC 和 Service 的创建：

```
# kubectl create -f skydns-rc.yaml
# kubectl create -f skydns-svc.yaml
```

查看 RC、Pod 和 Service，确保容器成功启动：

```
# kubectl get rc --namespace=kube-system
```

NAME	DESIRED	CURRENT	AGE
------	---------	---------	-----

```
kube-dns-v11          1          1          1d

# kubectl get pods --namespace=kube-system
NAME                  READY    STATUS    RESTARTS    AGE
kube-dns-v11-6dlwu    4/4     Running   0           1d

# kubectl get services --namespace=kube-system
NAME          CLUSTER-IP    EXTERNAL-IP    PORT(S)          AGE
kube-dns      169.169.0.100 <none>         53/UDP,53/TCP    1d
```

然后，我们为 redis-master 应用创建一个 Service。

### redis-master-service.yaml

```
apiVersion: v1
kind: Service
metadata:
  name: redis-master
  labels:
    name: redis-master
spec:
  ports:
    - port: 6379
      targetPort: 6379
  selector:
    name: redis-master
```

查看创建好的 redis-master service:

```
# kubectl get services
NAME          CLUSTER-IP    EXTERNAL-IP    PORT(S)          AGE
redis-master  169.169.8.10  <none>         6379/TCP         1h
```

可以看到，系统为 redis-master 服务分配了一个虚拟 IP 地址：169.169.8.10。

至此，在 Kubernetes 集群内的虚拟 DNS 服务就搭建好了。在需要访问 redis-master 的应用中，仅需要配置上 redis-master Service 的名称和服务的端口号，就能够访问到 redis-master 应用了，让我们回顾一下 redis-slave 应用需要访问 redis-master 的配置内容：

redis-slave 镜像的启动脚本/run.sh 的内容为：

```
if [[ ${GET_HOSTS_FROM:-dns} == "env" ]]; then
  redis-server --slaveof ${REDIS_MASTER_SERVICE_HOST} 6379
else
  redis-server --slaveof redis-master 6379
fi
```

在使用 DNS 模式的情况下，redis-slave 配置的 Master 地址为：redis-master:6379。通过服务名进行配置，能够极大地简化客户端应用对后端服务变化的感知，包括服务虚拟 IP 地址的变化、

服务后端 Pod 的变化等，对应用程序的微服务架构实现提供了强有力的支撑。

#### 4. 通过 DNS 查找 Service

接下来使用一个带有 nslookup 工具的 Pod 来验证 DNS 服务是否能够正常工作：

##### busybox.yaml

```
apiVersion: v1
kind: Pod
metadata:
  name: busybox
  namespace: default
spec:
  containers:
  - name: busybox
    image: gcr.io/google_containers/busybox
    command:
      - sleep
      - "3600"
```

运行 `kubectl create -f busybox.yaml` 完成创建。

在该容器成功启动后，通过 `kubectl exec <container_id> nslookup` 进行测试：

```
# kubectl exec busybox -- nslookup redis-master
Server:      169.169.0.100
Address 1: 169.169.0.100

Name:        redis-master
Address 1: 169.169.8.10
```

可以看到，通过 DNS 服务器 169.169.0.100 成功找到了名为“redis-master”服务的 IP 地址：169.169.8.10。

如果某个 Service 属于不同的命名空间，那么在进行 Service 查找时，需要带上 namespace 的名字。下面以查找 kube-dns 服务为例：

```
# kubectl exec busybox -- nslookup kube-dns.kube-system
Server:      169.169.0.100
Address 1: 169.169.0.100

Name:        kube-dns.kube-system
Address 1: 169.169.0.100
```

如果仅使用“kube-dns”来进行查找，则将会失败：

```
nslookup: can't resolve 'kube-dns'
```

## 5. DNS 服务的工作原理解析

让我们看看 DNS 服务背后的工作原理。

(1) kube2sky 容器应用通过调用 Kubernetes Master 的 API 获得集群中所有 Service 的信息，并持续监控新 Service 的生成，然后写入 etcd 中。

查看 etcd 中存储的 Service 信息：

```
# kubectl exec kube-dns-v8-5tpm2 -c etcd --namespace=kube-system etcdctl ls
/skydns/local/cluster
/skydns/local/cluster/default
/skydns/local/cluster/svc
/skydns/local/cluster/kube-system
```

可以看到在 skydns 键下面，根据我们配置的域名（cluster.local）生成了 local/cluster 子键，接下来是 namespace（default 和 kube-system）和 svc（下面也按 namespace 生成子键）。

查看 redis-master 服务对应的键值：

```
# kubectl exec kube-dns-v8-5tpm2 -c etcd --namespace=kube-system etcdctl get /
skydns/local/cluster/default/redis-master
{"host":"169.169.8.10","priority":10,"weight":10,"ttl":30,"targetstrip":0}
```

可以看到，redis-master 服务对应的完整域名为 redis-master.default.cluster.local，并且其 IP 地址为 169.169.8.10。

(2) 根据 kubelet 启动参数的设置（--cluster\_dns），kubelet 会在每个新创建的 Pod 中设置 DNS 域名解析配置文件/etc/resolv.conf 文件，在其中增加了一条 nameserver 配置和一条 search 配置：

```
nameserver 169.169.0.100
search default.svc.cluster.local svc.cluster.local cluster.local localdomain
```

通过名字服务器 169.169.0.100 访问的实际上就是 skydns 在 53 端口上提供的 DNS 解析服务。

(3) 最后，应用程序就能够像访问网站域名一样，仅仅通过服务的名字就能访问到服务了。

仍然以 redis-slave 为例，假设已经启动了 redis-slave Pod，登录 redis-slave 容器进行查看，可以看到其通过 DNS 域名服务找到了 redis-master 的 IP 地址 169.169.8.10，并成功建立了连接。

## 6. DNS 服务的演进

在后续的版本中，skydns 将被更新为 kubedns（不再使用 etcd 数据库），同时增加 DNS 专

用的 HPA 控制器，以自动扩展 DNS 容器的副本数量。但其工作机制没有发生变化，都是为了使客户端能够通过服务的名称访问到后端的服务。DNS 服务是 Kubernetes 集群中服务发现的最核心组件，建议将其作为标准配置，在安装集群时部署。

### 2.4.6 自定义 DNS 和上游 DNS 服务器

在实际环境中，很多用户都有自己的私有域名区域，并且希望能够集成到 Kubernetes DNS 的命名空间中，例如混合云用户可能希望能在集群内解析其内部的“.corp”域名；用户也可能已存在一个未被 Kubernetes 管理的服务发现系统（例如 Consul）来完成域名解析。从 Kubernetes v1.6 版本开始，用户可以在 Kubernetes 集群内配置私有 DNS 区域（通常称为存根域 Stub Domain）和外部的上游域名服务了，本节讲解如何使用这一功能。

#### 1. Kubernetes 默认的域名解析流程

Kubernetes 目前在 Pod 定义中支持两个 DNS 策略：Default 和 ClusterFirst，dnsPolicy 默认为 ClusterFirst。如果将 dnsPolicy 设置为 Default，域名解析配置则完全从 Pod 所在的节点（/etc/resolv.conf）继承而来，如图 2.15 所示。

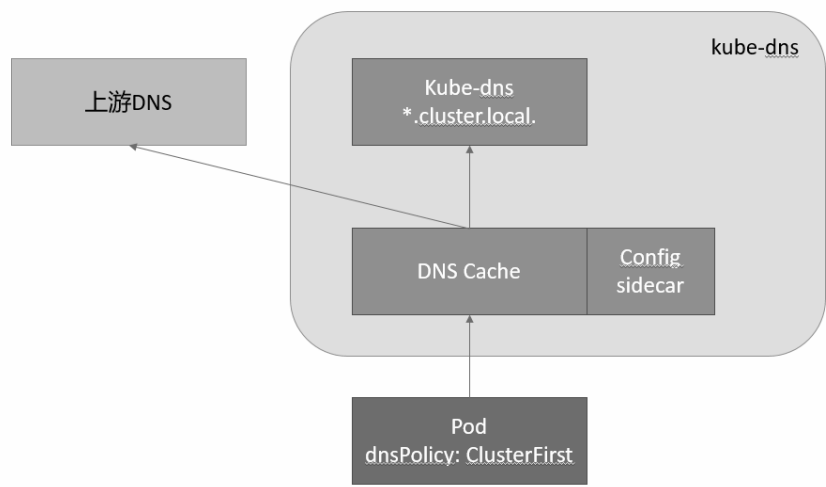


图 2.15 Kubernetes 默认的域名解析流程

如果将 dnsPolicy 设置为 ClusterFirst，则 DNS 查询会被发送到 kube-dns（skydns）服务。kube-dns 服务负责以集群域名为后缀（例如.cluster.local）进行服务名的解析。其他域名查询（例如 www.kubernetes.io）会被转发给节点上定义的上游域名服务器。

在这一功能推出之前，通常需要利用替换上游 DNS 为自定义解析的方式来完成存根域查询，但这样会使自定义域名解析器成为 DNS 解析过程中的一个高风险因素。本功能让用户无须对整个 DNS 路径进行改造就完成自定义 DNS 解析的过程。

## 2. 自定义 DNS 的方式

从 Kubernetes v1.6 版本开始，集群管理员可以使用 ConfigMap 指定自定义的存根域和上游 DNS Server。

下例中的配置包含一个存根域和两个上游域名服务器，对域名后缀为 out-of.kubernetes 的查询会被发送到地址为 10.140.0.5 的 DNS 服务，并设置 8.8.8.8 和 8.8.4.4（Google DNS 服务器）为上游 DNS 服务器地址。

```
apiVersion: v1
kind: ConfigMap
metadata:
  name: kube-dns
  namespace: kube-system
data:
  stubDomains: |
    {"out-of.kubernetes": ["10.140.0.5"]}
  upstreamNameservers: |
    ["8.8.8.8", "8.8.4.4"]
```

主要参数说明如下。

- ◎ **stubDomains (可选)**: JSON 格式的存根域定义，Key 为 DNS 后缀 (out-of.kubernetes)，值是一个 JSON 数组，表示一组 DNS 服务的 IP 地址。注意，目标域名服务器也可以是 Kubernetes 服务名，例如可以用 dnsmasq 把自定义 DNS 导出到 ClusterDNS 的命名空间中。
- ◎ **upstreamNameservers**: 一个 DNS IP 组成的 JSON 数组。注意，如果指定了这个值，那么从节点的域名服务设置 (/etc/resolv.conf) 继承过来的值就会被覆盖。本字段限制最多指定 3 个 IP 地址。

图 2.16 显示了配置中所指示的 DNS 域名解析流程。当 dnsPolicy 设置为 ClusterFirst 时，DNS 查询首先被发送到 kube-dns 的 DNS 缓存层。从这里开始检查域名后缀，然后发送到指定的 DNS。在本例中，集群后缀的域名 (.cluster.local) 被发送到 kube-dns，域名后缀符合配置 (.out-of.kubernetes) 的会被发送配置中的自定义解析服务器，不符合以上后缀的其他查询被转发到上游 DNS 进行解析。

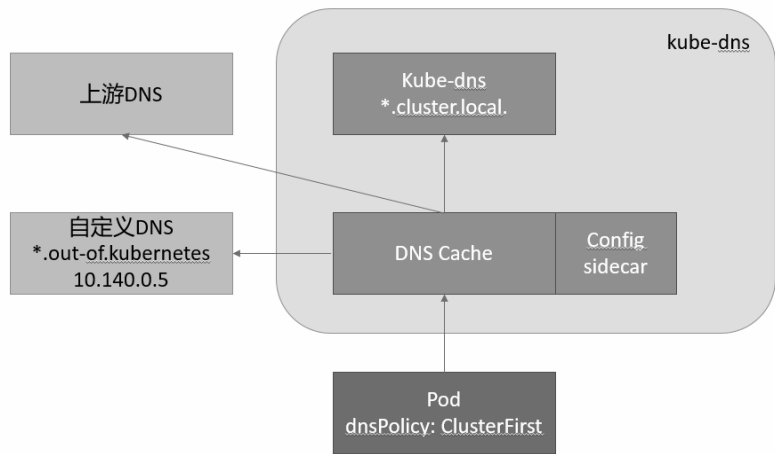


图 2.16 自定义 DNS 和上游 DNS 的域名解析流程

表 2.18 说明了域名解析的顺序。

表 2.18 域名解析的顺序

域 名	使用的 DNS 服务器
kubernetes.default.svc.cluster.local	kube-dns
foo.out-of.kubernetes	自定义 DNS: 10.140.0.5
widget.com	上游 DNS

下面通过一个例子说明如何在 Kubernetes 中使用自定义的 DNS 服务。

### 1) 安装 dnsmasq 作为自定义的 DNS 服务器

在节点 10.140.0.5 上安装一个 dnsmasq 服务器（yum install 或 apt-get install），并创建一条主机记录供 dnsmasq 使用。

生成一个自定义的 DNS 记录文件/tmp/hosts：

```
# echo "192.168.10.2 server.out-of.kubernetes" > /tmp/hosts
```

启动 DNS 服务：

```
# dnsmasq -q -d -h -q -R -H /tmp/hosts
```

参数说明如下。

- -d: 以 debug 模式启动，在前台运行，便于观察日志。
- -q: 输出查询记录。
- -h: 不使用/etc/hosts。



- ◎ **-R:** 不使用/etc/resolve.conf。
- ◎ **-H:** 使用自定义的文件作为 DNS 记录。

这样就启动了一个 DNS 服务器，可以直观地看到其工作状态：

```
dnsmasq: started, version 2.66 cachesize 150
dnsmasq: compile time options: IPv6 GNU-getopt DBus no-lln IDN DHCP DHCPv6 no-Lua
TFTP no-contrack ipset auth
dnsmasq: warning: no upstream servers configured
dnsmasq: read /tmp/hosts - 1 addresses
```

输出信息表明 dnsmasq 没有设置上游服务器，并且从文本文件中导入了一个主机记录。

## 2) 创建自定义 DNS 的 Configmap

配置文件 dns-configmap.yaml 的内容如下：

```
apiVersion: v1
kind: ConfigMap
metadata:
  name: kube-dns
  namespace: kube-system
data:
  stubDomains: |
    {"out-of.kubernetes": ["10.140.0.5"]}
  upstreamNameservers: |
    ["8.8.8.8", "8.8.4.4"]
```

使用 `kubectrl create` 命令完成创建：

```
# kubectrl create -f dns-configmap.yaml
configmap "kube-dns" created
```

## 3) 运行一个容器，进入容器内部查看自定义的 DNS 配置是否生效

```
apiVersion: v1
kind: Pod
metadata:
  name: tester
spec:
  dnsPolicy: ClusterFirst
  containers:
  - name: busybox
    image: busybox
    command: ["sleep"]
    args: ["3600"]
```

Pod 成功启动后，通过 `kubectrl exec -it tester -- sh` 命令进入 Pod。

在 Pod 内执行 `ping` 命令尝试解析域名 `server.out-of.kubernetes` 的 IP 地址：

```
# ping server.out-of.kubernetes
```

查看 dnsmasq 的输出日志，会看到如下内容：

```
.....
dnsmasq: query[A] server.out-of.kubernetes from 10.140.0.19
dnsmasq: /tmp/hosts server.out-of.kubernetes is 192.168.10.2
.....
```

说明对域名“server.out-of.kubernetes”的解析请求被指派给了自定义域名服务器 10.140.0.5（10.140.0.19 是 Pod 运行节点的 IP 地址），并成功解析了该域名的 IP 地址“192.168.10.2”。

## 2.4.7 Ingress：HTTP 7 层路由机制

根据前面对 Service 的使用说明，我们知道 Service 的表现形式为 IP:Port，即工作在 TCP/IP 层。而对于基于 HTTP 的服务来说，不同的 URL 地址经常对应到不同的后端服务或者虚拟服务器（Virtual Host），这些应用层的转发机制仅通过 Kubernetes 的 Service 机制是无法实现的。从 Kubernetes v1.1 版本开始新增 Ingress 资源对象，用于将不同 URL 的访问请求转发到后端不同的 Service，以实现 HTTP 层的业务路由机制。Kubernetes 使用一个 Ingress 策略定义和一个具体的 Ingress Controller，两者结合并实现了一个完整的 Ingress 负载均衡器。

使用 Ingress 进行负载分发时，Ingress Controller 将基于 Ingress 规则将客户端请求直接转发到 Service 对应的后端 Endpoint（即 Pod）上，这样会跳过 kube-proxy 的转发功能，kube-proxy 不再起作用。如果 Ingress Controller 提供的是对外服务，则实际上实现的是边缘路由器的功能。

图 2.17 显示了一个典型的 HTTP 层路由的例子。

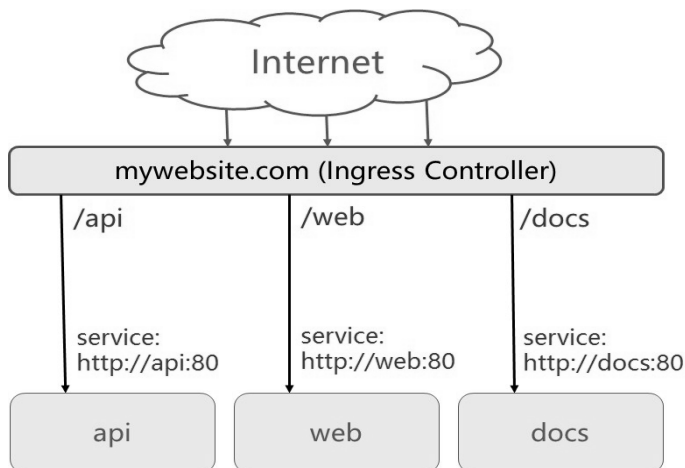


图 2.17 一个典型的 HTTP 层路由的例子

- ◎ 对 `http://mywebsite.com/api` 的访问将被路由到后端名为“api”的 Service。
- ◎ 对 `http://mywebsite.com/web` 的访问将被路由到后端名为“web”的 Service。
- ◎ 对 `http://mywebsite.com/doc` 的访问将被路由到后端名为“doc”的 Service。

为使用 Ingress，需要创建 Ingress Controller（带一个默认 backend 服务）和 Ingress 策略设置来共同完成。下面通过一个例子分三步说明 Ingress Controller 和 Ingress 策略的配置方法和客户端如何访问 Ingress 提供的服务。

## 1. 创建 Ingress Controller 和默认的 backend 服务

在定义 Ingress 策略之前，需要先部署 Ingress Controller，以实现为所有后端 Service 提供一个统一的入口。Ingress Controller 需要实现基于不同 HTTP URL 向后转发的负载分发规则，并可以灵活设置 7 层的负载分发策略。如果公有云服务商能够提供该类型的 HTTP 路由 LoadBalancer，则也可设置其为 Ingress Controller。

在 Kubernetes 中，Ingress Controller 将以 Pod 的形式运行，监控 apiserver 的/ingress 接口后端的 backend services，如果 service 发生变化，则 Ingress Controller 应自动更新其转发规则。

下面的例子使用 Nginx 来实现一个 Ingress Controller，需要实现的基本逻辑如下。

- (1) 监听 apiserver，获取全部 ingress 的定义。
- (2) 基于 ingress 的定义，生成 Nginx 所需的配置文件/etc/nginx/nginx.conf。
- (3) 执行 `nginx -s reload` 命令，重新加载 `nginx.conf` 配置文件的内容，

基于 Go 语言的核心代码实现如下：

```
for {
    rateLimiter.Accept()
    ingresses, err := ingClient.List(labels.Everything(), fields.Everything())
    if err != nil || reflect.DeepEqual(ingresses.Items, known.Items) {
        continue
    }
    if w, err := os.Create("/etc/nginx/nginx.conf"); err != nil {
        log.Fatalf("Failed to open %v: %v", nginxConf, err)
    } else if err := tpl.Execute(w, ingresses); err != nil {
        log.Fatalf("Failed to write template %v", err)
    }
    shellOut("nginx -s reload")
}
```

本例使用谷歌提供的 `nginx-ingress-controller` 镜像来创建 Ingress Controller。该 Ingress Controller 以 `daemonset` 的形式进行创建，在每个 Node 上都将启动一个 Nginx 服务。

这里为 Nginx 容器设置了 `hostPort`，将容器应用监听的 80 和 443 端口号映射到物理机上，使得客户端应用可以通过 URL 地址 “`http://物理机 IP:80`” 或 “`https://物理机 IP:443`” 来访问该 Ingress Controller。这使得 Nginx 类似于通过 NodePort 映射到物理机的 Service，成为代替 kube-proxy 的 HTTP 层的 Load Balancer。

#### **nginx-ingress-daemonset.yaml**

```
apiVersion: extensions/v1beta1
kind: DaemonSet
metadata:
  name: nginx-ingress-lb
  labels:
    name: nginx-ingress-lb
  namespace: kube-system
spec:
  template:
    metadata:
      labels:
        name: nginx-ingress-lb
    spec:
      terminationGracePeriodSeconds: 60
      containers:
        - image: gcr.io/google_containers/nginx-ingress-controller:0.9.0-beta.2
          name: nginx-ingress-lb
          readinessProbe:
            httpGet:
              path: /healthz
              port: 10254
              scheme: HTTP
          livenessProbe:
            httpGet:
              path: /healthz
              port: 10254
              scheme: HTTP
          initialDelaySeconds: 10
          timeoutSeconds: 1
          ports:
            - containerPort: 80
              hostPort: 80
            - containerPort: 443
              hostPort: 443
          env:
            - name: POD_NAME
              valueFrom:
                fieldRef:
                  fieldPath: metadata.name
```

```

- name: POD_NAMESPACE
  valueFrom:
    fieldRef:
      fieldPath: metadata.namespace
args:
- /nginx-ingress-controller
- --default-backend-service=$(POD_NAMESPACE)/default-http-backend

```

为了让 Ingress Controller 能够正常启动，还需要为它配置一个默认的 backend，用于在客户端访问的 URL 地址不存在时，能够返回一个正确的 404 应答。这个 backend 服务用任何应用实现都可以，只要满足默认对/路径的访问返回 404 应答，并且提供/healthz 路径以使 kubelet 完成对它的健康检查。另外，由于 Nginx 通过 default-backend-service 的服务名称（Service Name）去访问它，所以需要 DNS 服务正确运行。

#### **default-backend.yaml**

```

apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: default-http-backend
  labels:
    k8s-app: default-http-backend
  namespace: kube-system
spec:
  replicas: 1
  template:
    metadata:
      labels:
        k8s-app: default-http-backend
    spec:
      terminationGracePeriodSeconds: 60
      containers:
      - name: default-http-backend
        # Any image is permissable as long as:
        # 1. It serves a 404 page at /
        # 2. It serves 200 on a /healthz endpoint
        image: gcr.io/google_containers/defaultbackend:1.0
        livenessProbe:
          httpGet:
            path: /healthz
            port: 8080
            scheme: HTTP
          initialDelaySeconds: 30
          timeoutSeconds: 5
        ports:
        - containerPort: 8080
        resources:

```

```
      limits:
        cpu: 10m
        memory: 20Mi
      requests:
        cpu: 10m
        memory: 20Mi
---
apiVersion: v1
kind: Service
metadata:
  name: default-http-backend
  namespace: kube-system
  labels:
    k8s-app: default-http-backend
spec:
  ports:
    - port: 80
      targetPort: 8080
  selector:
    k8s-app: default-http-backend
```

通过 `kubectl create` 命令创建 backend 服务：

```
# kubectl create -f default-backend.yaml
deployment "default-http-backend" created
service "default-http-backend" created
```

创建 `nginx-ingress-controller`：

```
# kubectl create -f nginx-ingress-daemonset.yaml
daemonset "nginx-ingress-lb" created
```

查看 `default-http-backend` 和 `nginx-ingress-controller` 容器是否正确运行：

```
# kubectl get po --namespace=kube-system
```

NAME	READY	STATUS	RESTARTS	AGE
default-http-backend-1132503640-84lnv	1/1	Running	0	3m
kube-dns-v11-z3cb0	4/4	Running	0	10m
nginx-ingress-lb-5jbbv	1/1	Running	0	3m
nginx-ingress-lb-60j7h	1/1	Running	0	3m
nginx-ingress-lb-dttr9	1/1	Running	0	3m

用 `curl` 访问任意 Node 的 80 端口号，验证 `nginx-ingress-controller` 和 `default-http-backend` 服务正常工作：

```
# curl k8s-node-2
default backend - 404
```

## 2. 定义 Ingress 策略

本例对 mywebsite.com 网站的访问设置 Ingress 策略，定义对其/demo 路径的访问转发到后端 webapp Service 的规则：

```
apiVersion: extensions/v1beta1
kind: Ingress
metadata:
  name: mywebsite-ingress
spec:
  rules:
  - host: mywebsite.com
    http:
      paths:
      - path: /demo
        backend:
          serviceName: webapp
          servicePort: 8080
```

这个 Ingress 的定义，说明对目标地址 http://mywebsite.com/demo 的访问将被转发到集群中的 Service webapp 即 webapp:8080/demo 上。

在 Ingress 生效之前，需要先将 webapp 服务部署完成。同时需要注意 Ingress 中 path 的定义，需要与后端真实 Service 提供的 path 一致，否则将会转发到一个不存在的 path 上，引发错误。这里以第 1 章的例子为例，假设 myweb 服务已经部署完毕并正常运行，myweb 提供的 Web 服务的路径也为/demo。

创建该 Ingress：

```
# kubectl create -f ingress.yaml
ingress "mywebsite-ingress" created

# kubectl get ingress -o wide
NAME                                HOSTS                                ADDRESS                                PORTS
AGE
mywebsite-ingress  mywebsite.com  192.168.18.3,192.168.18.4,192.168.18.5
80                               59s
```

在成功创建该 Ingress 后，查看其 ADDRESS 列，如果显示了所有 nginx-ingress-controller Pod 的 IP 地址，则表示 Nginx 已经设置好后端 Service 的 Endpoint，该 Ingress 可以正常工作了。如果 ADDRESS 列为空，则通常说明 Nginx 未能正确连接到后端 Service，需要排查。

登录任一 nginx-ingress-controller Pod，查看其自动生成的 nginx.conf 配置文件内容，可以看到对 mywebsite.com/demo 的转发规则的正确配置：

```
daemon off;
worker_processes 1;
```

```
.....
http {
    real_ip_header    X-Forwarded-For;
    set_real_ip_from  0.0.0.0/0;
    real_ip_recursive on;
    .....
    upstream default-myweb-8080 {
        least_conn;
        server 172.17.1.5:8080 max_fails=0 fail_timeout=0;
    }
    .....
    server {
        server_name mywebsite.com;
        listen [::]:80;

        location /demo {
            set $proxy_upstream_name "default-myweb-8080";

            port_in_redirect off;
            client_max_body_size          "1m";

            proxy_set_header Host          $host;

            # Pass Real IP
            proxy_set_header X-Real-IP     $remote_addr;

            # Allow websocket connections
            proxy_set_header                Upgrade          $http_upgrade;
            proxy_set_header                Connection
$connection_upgrade;

            proxy_set_header X-Forwarded-For $proxy_add_x_forwarded_for;
            proxy_set_header X-Forwarded-Host $host;
            proxy_set_header X-Forwarded-Port $pass_port;
            proxy_set_header X-Forwarded-Proto $pass_access_scheme;

            .....

```

### 3. 客户端访问 <http://mywebsite.com/demo>

由于 Ingress Controller 容器通过 hostPort 将服务端口号 80 映射到了所有 Node 上，所以客户端可以通过任意 Node 访问 mywebsite.com 提供的服务。

需要说明的是，客户端只能通过域名“mywebsite.com”访问服务，这时要求客户端或者 DNS 能够将 mywebsite.com 域名解析到后端多个 Node 的真实 IP 地址上。

通过 curl 访问 mywebsite.com 提供的服务（可以用--resolve 参数模拟 DNS 解析，目标地址



为域名；也可以用-H 'Host:mywebsite.com'参数设置 HTTP 头中要访问的域名，目标地址为 IP 地址），可以得到 myweb 服务返回的网页内容。

```
# curl --resolve mywebsite.com:80:192.168.18.3 http://mywebsite.com/demo/

或

# curl -H 'Host:mywebsite.com' http://192.168.18.3/demo/
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=utf-8">
<title>HPE University Docker&Kubernetes Learning</title>
</head>
<body align="center">

    <h2>Congratulations!!</h2>
    <br></br>
    <input type="button" value="Add..."
onclick="location.href='input.html'" >
    <br></br>
    <TABLE align="center" border="1" width="600px">
    <TR>
    <TD>Name</TD>
    <TD>Level(Score)</TD>
    </TR>

    <TR>
    <TD>google</TD>
    <TD>100</TD>
    </TR>

    <TR>
    <TD>docker</TD>
    <TD>100</TD>
    </TR>

    <TR>
    <TD>teacher</TD>
    <TD>100</TD>
    </TR>

    <TR>
    <TD>HPE</TD>
    <TD>100</TD>
    </TR>
```

```
<TR>
  <TD>our team</TD>
  <TD>100</TD>
</TR>

<TR>
  <TD>me</TD>
  <TD>100</TD>
</TR>

</TABLE>

</body>
</html>
```

如果是通过浏览器访问，那么需要先在本地上设置域名 mywebsite.com 对应的 IP 地址，再到浏览器上进行访问。以 Windows 为例，修改 C:\Windows\System32\drivers\etc\hosts 文件，加入一行记录：

```
192.168.18.3 mywebsite.com
```

然后在浏览器地址栏输入 http://mywebsite.com/demo/，就能够访问 Ingress 提供的服务了，如图 2.18 显示。

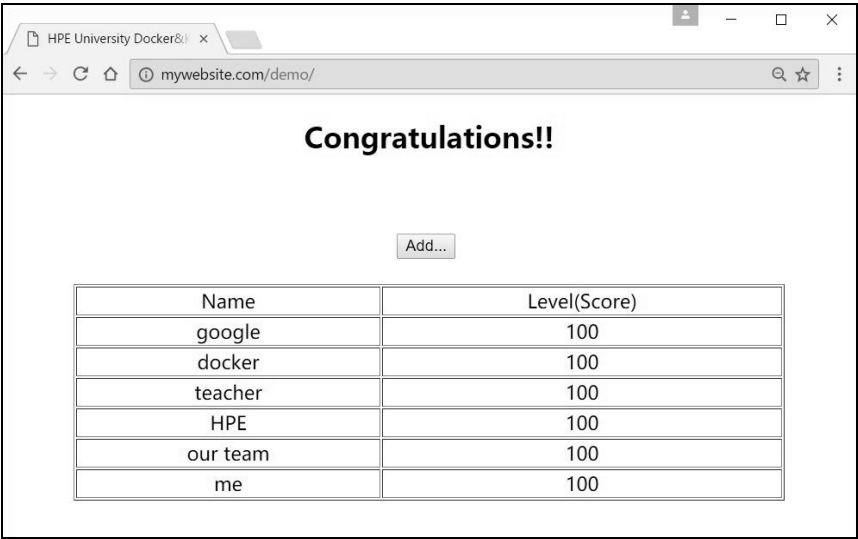


图 2.18 浏览器访问 Ingress 服务

#### 4. Ingress 的策略配置技巧

为了实现灵活的负载分发策略，Ingress 策略可以按多种方式进行配置，下面对几种常见的 Ingress 转发策略进行说明。

##### 1) 转发到单个后端服务上

基于这种设置，客户端到 Ingress Controller 的访问请求都将被转发到后端的唯一 Service 上。这种情况下 Ingress 无须定义任何 rule。

通过如下所示的设置，对 Ingress Controller 的访问请求都将被转发到 “myweb:8080” 这个服务上。

```
apiVersion: extensions/v1beta1
kind: Ingress
metadata:
  name: test-ingress
spec:
  backend:
    serviceName: myweb
    servicePort: 8080
```

##### 2) 同一域名下，不同的 URL 路径被转发到不同的服务上

这种配置常用于一个网站通过不同的路径提供不同的服务的场景，例如/web 表示访问 Web 页面，/api 表示访问 API 接口，对应到后端的两个服务，通过 Ingress 的设置很容易就能将基于 URL 路径的转发规则定义出来。

通过如下所示的设置，对 “mywebsite.com/web” 的访问请求将被转发到 “web-service:80” 服务上；对 “mywebsite.com/api” 的访问请求将被转发到 “api-service:80” 服务上。

```
apiVersion: extensions/v1beta1
kind: Ingress
metadata:
  name: test-ingress
spec:
  rules:
  - host: mywebsite.com
    http:
      paths:
      - path: /web
        backend:
          serviceName: web-service
          servicePort: 80
      - path: /api
        backend:
          serviceName: api-service
          servicePort: 8081
```

### 3) 不同的域名（虚拟主机名）被转发到不同的服务上

这种配置常用于一个网站通过不同的域名或虚拟主机名提供不同的服务的场景，例如 foo.bar.com 域名由 service1 提供服务，bar.foo.com 域名由 service2 提供服务。

通过如下所示的设置，对“foo.bar.com”的访问请求将被转发到“service1:80”服务上，对“bar.foo.com”的访问请求将被转发到“service2:80”服务上：

```
apiVersion: extensions/v1beta1
kind: Ingress
metadata:
  name: test
spec:
  rules:
    - host: foo.bar.com
      http:
        paths:
          - backend:
              serviceName: service1
              servicePort: 80
    - host: bar.foo.com
      http:
        paths:
          - backend:
              serviceName: service2
              servicePort: 80
```

### 4) 不使用域名的转发规则

这种配置用于一个网站不使用域名直接提供服务的场景，此时通过任意一台运行 ingress-controller 的 Node 都能访问到后端的服务。

以上节的后端服务 webapp 为例，下面的配置为将“<ingress-controller-ip>/demo”的访问请求转发到“webapp:8080/demo”服务上。

```
apiVersion: extensions/v1beta1
kind: Ingress
metadata:
  name: test-ingress
spec:
  rules:
    - http:
        paths:
          - path: /demo
            backend:
              serviceName: webapp
              servicePort: 8080
```

注意，使用无域名的 Ingress 转发规则时，将默认禁用非安全 HTTP，强制启用 HTTPS。例如，当使用 Nginx 作为 Ingress Controller 时，其配置文件/etc/nginx/nginx.conf 中将会自动设置下面的规则，将全部 HTTP 的访问请求直接返回 301 错误。

```
.....
# enforce ssl on server side
if ($pass_access_scheme = http) {
    return 301 https://$best_http_host$request_uri;
}
.....
```

客户端使用 HTTP 访问将得到 301 的错误应答：

```
# curl http://192.168.18.3/demo/
<html>
<head><title>301 Moved Permanently</title></head>
<body bgcolor="white">
<center><h1>301 Moved Permanently</h1></center>
<hr><center>nginx/1.13.0</center>
</body>
</html>
```

使用 HTTPS 能够访问成功：

```
# curl -k https://192.168.18.3/demo/
.....
    <h2>Congratulations!!</h2>
    <br></br>
    <input type="button" value="Add..."
onclick="location.href='input.html'" >
    <br></br>
    <TABLE align="center" border="1" width="600px">
    <TR>
    <TD>Name</TD>
    <TD>Level(Score)</TD>
    </TR>
.....
```

可以在 Ingress 的定义中设置一个 annotation “ingress.kubernetes.io/ssl-redirect=false” 来关闭强制启用 HTTPS 的设置：

```
apiVersion: extensions/v1beta1
kind: Ingress
metadata:
  name: test-ingress
  annotations:
    ingress.kubernetes.io/ssl-redirect: "false"
spec:
  rules:
```

```
- http:
  paths:
  - path: /
    backend:
      serviceName: web-service
      servicePort: 8080
```

这样，到 Ingress Controller 的访问就可以使用 HTTP 了：

```
# curl http://192.168.18.3/demo/
.....
<h2>Congratulations!!</h2>
<br></br>
  <input type="button" value="Add..."
onclick="location.href='input.html'" >
    <br></br>
    <TABLE align="center" border="1" width="600px">
<TR>
  <TD>Name</TD>
  <TD>Level(Score)</TD>
</TR>
.....
```

## 5. Ingress 的 TLS 安全设置

为了使得 Ingress 提供 HTTPS 的安全访问，可以为 Ingress 中的域名进行 TLS 安全证书的设置。设置的步骤如下。

- (1) 创建自签名的密钥和 SSL 证书文件。
- (2) 将证书保存到 Kubernetes 中的一个 Secret 资源对象上。
- (3) 将该 Secret 对象设置到 Ingress 中。

根据提供服务的网站域名是一个还是多个，可以使用不同的操作完成前两步 SSL 证书和 Secret 对象的创建，在只有一个域名的情况下设置相对简单。第 3 步对于这两种场景来说是相同的。

对于只有一个域名的场景来说，可以通过 OpenSSL 工具直接生成密钥和证书文件，将命令行参数-subj 中的/CN 设置为网站域名：

```
# openssl req -x509 -nodes -days 5000 -newkey rsa:2048 -keyout tls.key -out tls.crt
-subj "/CN=mywebsite.com"
Generating a 2048 bit RSA private key
.....+++
.....+++
writing new private key to 'tls.key'
-----
```

上述命令将生成 `tls.key` 和 `tls.crt` 两个文件。

然后根据 `tls.key` 和 `tls.crt` 文件创建 `secret` 资源对象，有以下两种方法。

方法一：通过 `kubectl create secret tls` 命令直接通过 `tls.key` 和 `tls.crt` 文件创建 `secret` 对象。

```
# kubectl create secret tls mywebsite-ingress-secret --key tls.key --cert tls.crt
secret "mywebsite-ingress-secret" created
```

方法二：编辑 `mywebsite-ingress-secret.yaml` 文件，将 `tls.key` 和 `tls.crt` 文件的内容复制进去，使用 `kubectl create` 命令进行创建。

```
mywebsite-ingress-secret.yaml
apiVersion: v1
kind: Secret
metadata:
  name: mywebsite-ingress-secret
type: kubernetes.io/tls
data:
  tls.crt:
MIIDAzCCAeugAwIBAgIJALrTg9VLmFgdMA0GCSqGSIb3DQEBCwUAMBgxFjAUBgNVBAMMDW15d2Vic2l0
ZS5jb20wHhcNMTCwNDIzMTMwMjA1WhcNMzAxMjMxMTMwMjA1WjAYMRywFAYDVQQDDA1teXdlYnNpdGUu
Y29tMIIBIjANBgkqhkiG9w0BAQEFAAOCAQ8AMIIBCgKCAQEApl1y1rqlI3EQ5E0PjzW8Lc3heW4WYTyk
POisDT9Zgyc+TLPgj/YF4QnAuoIUaUNtXPlmINKuD9Fxmzh6q0oSBVb42BU0RzOTtvaCVOU+uoJ9MgJp
d7Bao5higTZMyvj5a1M9iwb7k4xRAsuGCh/jD08fj6tgJW4Wfzaw05w1pDd2fFDxYn34Malpg0xFebVa
iqBu9FL0JbiEimsV9y7V+g6jjfGffu2xl06X3svqAdfGhvs+uCTArAXiZgs279selXp834CG0MJeP7ta
mD44IfA2wkkmD+uCVjSEcNFsveY5cJevjf0PSE9g5wohSXphd1sIGyjEy2ApeIJBp8bQ+wIDAQABo1Aw
TjAdBgNVHQ4EFgQUjmpxpmdFPKwkr+A2XLF7oqro2GkwHwYDVR0jBBgwFoAUjmpxpmdFPKwkr+A2XLF7
oqro2GkwDAYDVR0TBABUwAwEB/zANBgkqhkiG9w0BAQsFAAOCAQEAAVXPYfagP1AIov3kXRhI3WfyCOIN
/sgNSqKM3FuykboSBN6clw4UhrpF71Hd4nt0myeyX/o69o20c9a9dIS2FEGKvfxZQ4sa99iI3qjoMAuu
f/Q9fDYIZ+k0YvY4pbcCqqOyICFBCMLlAct/aB0K1GBvC5k06vD4Rn2fOdVmKlOW+Zf41cxVIRZe/tQG
nZoEhtM6FQADrv1+jM5gjIKRX3s2/Jcxy5g2XLPqtSpzYA0F7FJyuFJXEG+P9X466xPi9ialUri66vkb
UVT6uLXGhhunsu6bZ/qwsm2HzdPo4WRQ3z2VhgFzHEzHVvX+CEyZ8fJGoSi7njapHb081RiztQ==
  tls.key:
MIIEvQIBADANBgkqhkiG9w0BAQEFAASCBKcwggSjAgEAAoIBAQCkvXLWurUjcrdKtQ+PNbwtzeF5bhZh
PKQ86KwNP1mDjz5Ms8aP9gXhCcC6ghQBQ21c+WYg0q4P0XHOaHqrShIFVvYjYFTRHM5029oJU5T66gn0y
Aml3sFqjmGKBNkzK+PlrUz2LBvuTjFECy4YKH+MM7x+Pq2AlbhZ/NrA7nDWkN3Z8UPFiffgxrWmDTEV5
tVqKoG70UvQ1uISKaxX3LtX6DqON8Z9+7bGXTpfey+oB18ag9L64JMCsBeJmBLbv2x7VenzfgIbQw14/
ulqYPjgh8DbCSSYP64JWNIRw0Wy95j1w16+N/Q9IT2DnCiFJemF3WwgbKMTLYA94gke/xtD7AgMBAAEC
ggEAFtNePqlRgvwYgzPX29YVFsOiAV28bDh8sW/SWBrRU9002uDtwSx7EmUNbyiA/bwJ8KdRlXr7uFG
B3gLa876pNmhQLdcqspKClUmiuUCKIJ7lZWIEt4aXStqae8BzEiWpwhnqhYxgD312sQ50jQII9mkFTUT
xbLBu1F95kxYjX2MfTrrvwroDLZEHCpcBY9hNUFhZaCdBBYKADmWo9eV/xZJ97ZAFpbpWyONrFjNwMj
jqCmxMx3HwOI/tLbhpvob6RT1UG1QUPlbB8aXR1FeSgt0NYhYwWKF7JSXcYBiYqubtd3T6RbTnJfK4b/
zuEUhdFN1lKJLcsVDVQZgMs04QKBgQDaJXaQ4hMKPH3CKdieAialj4rVAPyrAFYDMokW+7buZAgZ01a
rRtqFWLTtp6hwHqwTySHFYiRsK2Ikfct1H16hRn6FXbiPrFDP8gpYveu31Cd1qqYUYI7xaodWUiLldrt
eun9sLr3YYR7kaXYRenWZFjZbbUkq3KJfoh+uArPwwKBgQDA95Y4xhcL0F5pE/TLedj33WjRXMKXMXCHX
Gl3fTnBImorF7jF9e5fRK/v4YIHAMCon+6drwMv9KHFL0nvxPbgBECW1F2OfzmNgm617jkpcscQOVtuu
```

```
1+4gK+B2geQYRA2LhBk+9MtGQFmwSPgwSg+VHUrM28qhzUmTCN1etdpeaQKBgGAFqHSO44Kp1S8Lp6q0
kzpGeN7hEiIngaLh/y1j5pmTceFptocSa2sOf186azPyF3WDMC9SU3a/Q18vkoRGSeMcu6804y7AEK3V
RiI4402nvAm9GTLXDPsp+3XtllwNuSSBznCxx1ONOuH3uf/tp7GUYR0WgHHeCfKy71GNluJ1AoGAKhHQ
XnBRdfHno2EGbX9mniNXRs3DyZpkxlCpRpYDRNDRKz7y6ziW0LOWK4BezWLPwz/KMGPIFVlL2gv5mY6r
JLtQfTqsLZsBb36AZL+Q1sRQGBA3tNa+w6TNOwj2gZPUoCYcmu0jpB1DcHt4II8E9q18NviUJNJsx/GW
0Z80DIECgYEAXzQBh/ckRvRapRNOv8w9GRq3wTYyD9y15U+3ecEIzrrlg9bLOi/rktXy3vqL6kj6CFlp
wwRVLfjR83ulQPy3MpJNXyR1Bua+/FVn2xKwyYDuXaqs0vW3xLONVO7z44gAKmEQyDq2sir+vpayU4ps
fXXK06uifz6ELfVyY6XZvRA=
```

```
# kubectl create -f mywebsite-ingress-secret.yaml
secret "mywebsite-ingress-secret" created
```

如果提供服务的网站不止一个域名，例如前面第 3 种 Ingress 策略配置方式，则 SSL 证书需要使用额外的一个 x509 v3 配置文件辅助完成，在[alt\_names]段中完成多个 DNS 域名的设置。

编写 openssl.cnf 文件，内容为：

```
[req]
req_extensions = v3_req
distinguished_name = req_distinguished_name
[req_distinguished_name]
[ v3_req ]
basicConstraints = CA:FALSE
keyUsage = nonRepudiation, digitalSignature, keyEncipherment
subjectAltName = @alt_names
[alt_names]
DNS.1 = mywebsite.com
DNS.2 = mywebsite2.com
```

然后使用 OpenSSL 工具完成密钥和证书的创建。

首先生成自签名 CA 证书：

```
# openssl genrsa -out ca.key 2048
Generating RSA private key, 2048 bit long modulus
..+++
.....+++
e is 65537 (0x10001)
# openssl req -x509 -new -nodes -key ca.key -days 5000 -out ca.crt -subj
"/CN=mywebsite.com"
```

基于 openssl.cnf 和 ca 证书生成 ingress SSL 证书：

```
# openssl genrsa -out ingress.key 2048
Generating RSA private key, 2048 bit long modulus
.....+++
.....+++
e is 65537 (0x10001)
```



```
# openssl req -new -key ingress.key -out ingress.csr -subj "/CN=mywebsite.com"
-config openssl.cnf
# openssl x509 -req -in ingress.csr -CA ca.crt -CAkey ca.key -CAcreateserial -out
ingress.crt -days 5000 -extensions v3_req -extfile openssl.cnf
Signature ok
subject=/CN=mywebsite.com
Getting CA Private Key
```

然后根据 ingress.key 和 ingress.crt 文件创建 secret 资源对象，同样可以通过 kubectl create secret tls 命令或 yaml 配置文件生成。这里通过命令行直接生成：

```
# kubectl create secret tls mywebsite-ingress-secret --key ingress.key --cert
ingress.crt
secret "mywebsite-ingress-secret" created
```

至此，Ingress 的 TLS 证书就成功创建到 Secret 对象中了。

下面创建 Ingress 对象，在 tls 段引用刚刚创建好的 Secret 对象：

```
mywebsite-ingress-tls.yaml
apiVersion: extensions/v1beta1
kind: Ingress
metadata:
  name: mywebsite-ingress-tls
spec:
  tls:
    - hosts:
        - mywebsite.com
      secretName: mywebsite-ingress-secret
  rules:
    - host: mywebsite.com
      http:
        paths:
          - path: /demo
            backend:
              serviceName: myweb
              servicePort: 8080
```

之后，就可以通过 HTTPS 来访问 mywebsite.com 了。

以 curl 为例，访问 https://192.168.18.3/demo/：

```
# curl -H 'Host:mywebsite.com' -k https://192.168.18.3/demo/
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">
<html>
.....
  <h2>Congratulations!!</h2>
  <br></br>
  <input type="button" value="Add..."
```

```
onclick="location.href='input.html'" >
    <br></br>
    <TABLE align="center" border="1" width="600px">
    <TR>
    <TD>Name</TD>
    <TD>Level(Score)</TD>
    </TR>

    <TR>
    <TD>google</TD>
    <TD>100</TD>
    </TR>
    .....
</html>
```

如果是通过浏览器访问，则在浏览器的地址栏输入 `https://mywebsite.com/demo/` 来访问 Ingress 提供的服务，浏览器会提示不安全，如图 2.19 所示。

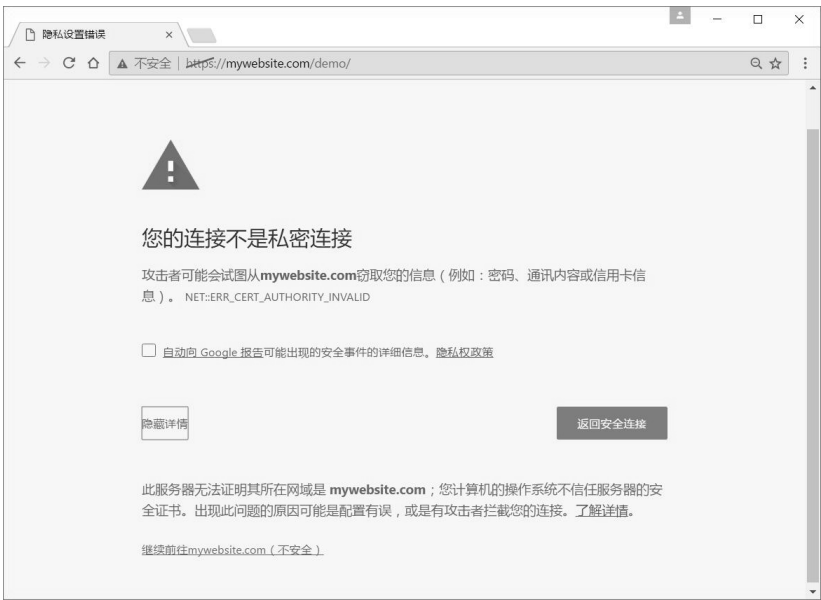


图 2.19 浏览器访问 Ingress HTTPS 服务的警告信息

单击“继续前往 mywebsite.com（不安全）”，访问后可看到 Ingress 后端服务提供的页面，如图 2.20 所示。

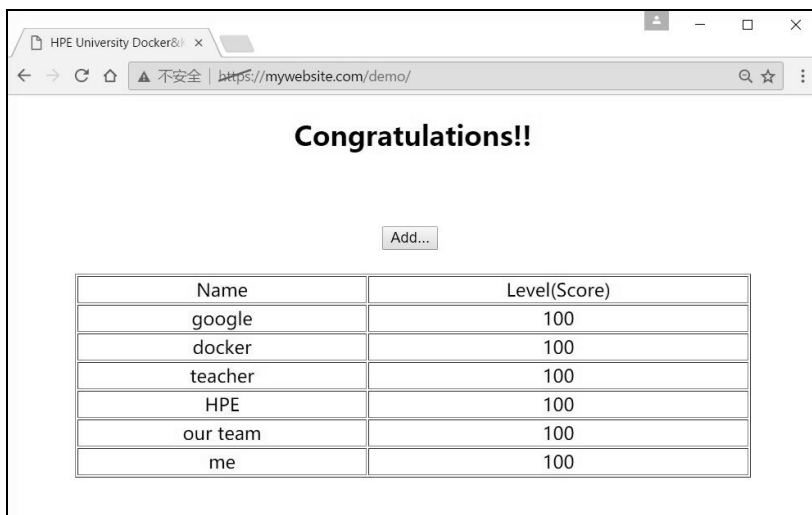


图 2.20 浏览器访问 Ingress HTTPS 服务

## 6. Ingress 的发展路线

当前的 Ingress 还是 Beta 版本，在 Kubernetes 的后续版本中将增加以下功能。

- ◎ 支持更多 TLS 选项，例如 SNI、重加密等。
- ◎ 支持 L4 和 L7 负载均衡策略（目前只支持 HTTP 层的规则）。
- ◎ 支持更多类型的 Ingress Controller，除了 Nginx，还有 HAProxy、Linkerd、traefik、AWS Application Load Balancer Ingress Controller 等已经实现了的 Ingress Controller，详情可参考 <https://github.com/kubernetes/ingress/blob/master/docs/catalog.md> 的说明。

# 第 3 章

## Kubernetes 核心原理

---

本章对 Kubernetes 的核心原理进行深入分析，首先从 API Server 的访问开始讲起，然后分析 Master 节点上 Controller Manager 各个组件的功能实现，以及 Scheduler 预选算法和优选算法。接下来讲解 Node 节点上的 kubelet 和 kube-proxy 组件的运行机制。最后，深入分析安全机制和网络原理。

### 3.1 Kubernetes API Server 原理分析

---

总体来看，Kubernetes API Server 的核心功能是提供了 Kubernetes 各类资源对象（如 Pod、RC、Service 等）的增、删、改、查及 Watch 等 HTTP Rest 接口，成为集群内各个功能模块之间数据交互和通信的中心枢纽，是整个系统的数据总线 and 数据中心。除此之外，它还有以下一些功能特性。

- （1）是集群管理的 API 入口。
- （2）是资源配额控制的入口。
- （3）提供了完备的集群安全机制。

#### 3.1.1 Kubernetes API Server 概述

---

Kubernetes API Server 通过一个名为 kube-apiserver 的进程提供服务，该进程运行在 Master 节点上。在默认情况下，kube-apiserver 进程在本机的 8080 端口（对应参数--insecure-port）提

供 REST 服务。我们可以同时启动 HTTPS 安全端口（--secure-port=6443）来启动安全机制，加强 REST API 访问的安全性。

通常我们可以通过命令行工具 `kubectl` 来与 Kubernetes API Server 交互，它们之间的接口是 REST 调用。为了测试和学习 Kubernetes API Server 所提供的接口，我们也可以使用 `curl` 命令行工具进行快速验证。

比如，我们登录 Master 节点，运行下面的 `curl` 命令，得到以 JSON 方式返回的 Kubernetes API 的版本信息：

```
# curl localhost:8080/api
{
  "kind": "APIVersions",
  "versions": [
    "v1"
  ],
  "serverAddressByClientCIDRs": [
    {
      "clientCIDR": "0.0.0.0/0",
      "serverAddress": "192.168.18.131:6443"
    }
  ]
}
```

可以运行下面的命令，来查看 Kubernetes API Server 目前所支持的资源对象的种类：

```
# curl localhost:8080/api/v1
```

根据以上命令的输出，我们可以运行下面的 `curl` 命令，分别返回集群中的 Pod 列表、Service 列表、RC 列表等：

```
# curl localhost:8080/api/v1/pods
# curl localhost:8080/api/v1/services
# curl localhost:8080/api/v1/replicationcontrollers
```

如果我们只想对外暴露部分 REST 服务，则可以在 Master 或其他任何节点上通过运行 `kubectl proxy` 进程启动一个内部代理来实现。

运行下面的命令，我们在 8001 端口启动代理，并且拒绝客户端访问 RC 的 API：

```
# kubectl proxy --reject-paths="^/api/v1/replicationcontrollers" --port=8001
--v=2
Starting to serve on 127.0.0.1:8001
```

运行下面的命令进行验证：

```
# curl localhost:8001/api/v1/replicationcontrollers
<h3>Unauthorized</h3>
```

`kubectl proxy` 具有很多特性，最实用的一个特性是提供简单有效的安全机制，比如采用白

名单来限制非法客户端访问时，只要增加下面这个参数即可：

```
--accept-hosts="^localhost$,^127\\.0\\.0\\.1$,^\\[::1\\]$"

```

最后一种方式是通过编程的方式调用 Kubernetes API Server。具体使用场景又细分为以下两种。

第 1 种使用场景：运行在 Pod 里的用户进程调用 Kubernetes API，通常用来实现分布式集群搭建的目标。比如下面这段来自谷歌官方的 Elastic Search 集群例子中的代码，Pod 在启动的过程中通过访问 Endpoints 的 API，找到属于 elasticsearch-logging 这个 Service 的所有 Pod 副本的 IP 地址，用来构建集群，如图 3.1 所示。



图 3.1 应用程序编程访问 API Server

在上述使用场景中，Pod 中的进程如何知道 API Server 的访问地址呢？答案很简单，因为 Kubernetes API Server 本身也是一个 Service，它的名字就是“kubernetes”，并且它的 Cluster IP 地址是 Cluster IP 地址池里的第 1 个地址！另外，它所服务的端口是 HTTPS 端口 443，通过 `kubectl get service` 命令可以确认这一点：

```
# kubectl get service
NAME          CLUSTER-IP      EXTERNAL-IP      PORT(S)          AGE
kubernetes    169.169.0.1     <none>           443/TCP          30d

```

第 2 种使用场景：开发基于 Kubernetes 的管理平台。比如调用 Kubernetes API 来完成 Pod、Service、RC 等资源对象的图形化创建和管理界面，此时可以使用 Kubernetes 及各开源社区为开发人员提供的各种语言版本的 Client Library。我们会在后面介绍通过编程方式访问 API Server 的一些细节技术。

### 3.1.2 独特的 Kubernetes Proxy API 接口

前面我们说过，Kubernetes API Server 最主要的 REST 接口是资源对象的增、删、改、查，除此之外，它还提供了一类很特殊的 REST 接口——Kubernetes Proxy API 接口，这类接口的作用是代理 REST 请求，即 Kubernetes API Server 把收到的 REST 请求转发到某个 Node 上的 kubelet 守护进程的 REST 端口上，由该 kubelet 进程负责响应。

首先，我们来说说 Kubernetes Proxy API 里关于 Node 的相关接口，该接口的 REST 路径为 `/api/v1/proxy/nodes/{name}`，其中 `{name}` 为节点的名称或 IP 地址，包括以下几个具体接口：

```
/api/v1/proxy/nodes/{name}/pods/    #列出指定节点内所有 Pod 的信息
/api/v1/proxy/nodes/{name}/stats/    #列出指定节点内物理资源的统计信息
/api/v1/proxy/nodes/{name}/spec/     #列出指定节点的概要信息
```

例如当前 Node 节点的名字为 `k8s-node-1`，用下面的命令即可获取该节点上所有运行中的 Pod：

```
# curl localhost:8080/api/v1/proxy/nodes/k8s-node-1/pods
```

需要说明的是：这里获取的 Pod 的信息数据来自 Node 而非 etcd 数据库，所以两者可能在某些时间点会有偏差。此外，如果 kubelet 进程在启动时包含 `--enable-debugging-handlers=true` 参数，那么 Kubernetes Proxy API 还会增加下面的接口：

```
/api/v1/proxy/nodes/{name}/run        #在节点上运行某个容器
/api/v1/proxy/nodes/{name}/exec       #在节点上的某个容器中运行某条命令
/api/v1/proxy/nodes/{name}/attach     #在节点上 attach 某个容器
/api/v1/proxy/nodes/{name}/portForward #实现节点上的 Pod 端口转发
/api/v1/proxy/nodes/{name}/logs       #列出节点的各类日志信息，例如 tallylog、lastlog、
                                       #wtmp、ppp/、rsh/、audit/、tuned/和 anaconda/等
/api/v1/proxy/nodes/{name}/metrics    #列出和该节点相关的 Metrics 信息
/api/v1/proxy/nodes/{name}/runningpods #列出节点内运行中的 Pod 信息
/api/v1/proxy/nodes/{name}/debug/pprof #列出节点内当前 Web 服务的状态
                                       #包括 CPU 占用情况和内存使用情况等
```

接下来，我们来说说 Kubernetes Proxy API 里关于 Pod 的相关接口，通过这些接口，我们可以访问 Pod 里某个容器提供的服务（如 Tomcat 在 8080 端口的服务）：

```
/api/v1/proxy/namespaces/{namespace}/pods/{name}/{path:*} #访问 Pod 的某个服务接口
/api/v1/proxy/namespaces/{namespace}/pods/{name}          #访问 Pod
/api/v1/namespaces/{namespace}/pods/{name}/proxy/{path:*} #访问 Pod 的某个服务接口
/api/v1/namespaces/{namespace}/pods/{name}/proxy          #访问 Pod
```

在上面的 4 个接口里，后面两个接口的功能与前面两个完全一样，只是写法不同。下面我们用第 1 章的 Java Web 例子中的 Tomcat Pod 来说明上述 Proxy 接口的用法。

首先，得到 Pod 的名字：

```
# kubectl get pods
NAME                READY    STATUS    RESTARTS   AGE
mysql-c95jc         1/1     Running   0           8d
myweb-g9pmm         1/1     Running   0           8d
```

然后，运行下面的命令，会输出 Tomcat 的首页，即相当于访问 `http://localhost:8080/`：

```
# curl http://localhost:8080/api/v1/proxy/namespaces/default/pods/myweb-g9pmm/
```

我们也可以在浏览器中访问上面的地址，比如 Master 节点的 IP 地址是 192.168.18.131，我们在浏览器中输入 `http://192.168.18.131:8080/api/v1/proxy/namespaces/default/pods/myweb-g9pmm/`，就能够访问 Tomcat 首页了；而如果输入 `/api/v1/proxy/namespaces/default/pods/myweb-g9pmm/demo`，就能访问 Tomcat 中 Demo 应用的页面了。

看到这里，你可能明白 Pod 的 Proxy 接口的作用和意义了：在 Kubernetes 集群之外访问某个 Pod 容器的服务（HTTP 服务）时，可以用 Proxy API 实现，这种场景多用于管理目的，比如逐一排查 Service 的 Pod 副本，检查哪些 Pod 的服务存在异常问题。

最后我们说说 Service，Kubernetes Proxy API 也有 Service 的 Proxy 接口，其接口定义与 Pod 的接口定义基本一样：`/api/v1/proxy/namespaces/{namespace}/services/{name}`。比如，若我们想访问 myweb 这个 Service，则可以在浏览器里输入 `http://192.168.18.131:8080/api/v1/proxy/namespaces/default/services/myweb/demo/`。

### 3.1.3 集群功能模块之间的通信

从图 3.2 中可以看出，Kubernetes API Server 作为集群的核心，负责集群各功能模块之间的通信。集群内的各个功能模块通过 API Server 将信息存入 etcd，当需要获取和操作这些数据时，则通过 API Server 提供的 REST 接口（用 GET、LIST 或 WATCH 方法）来实现，从而实现各模块之间的信息交互。

常见的一个交互场景是 kubelet 进程与 API Server 的交互。每个 Node 节点上的 kubelet 每隔一个时间周期，就会调用一次 API Server 的 REST 接口报告自身状态，API Server 接收到这些信息后，将节点状态信息更新到 etcd 中。此外，kubelet 也通过 API Server 的 Watch 接口监听 Pod 信息，如果监听到新的 Pod 副本被调度绑定到本节点，则执行 Pod 对应的容器的创建和启动逻辑；如果监听到 Pod 对象被删除，则删除本节点上的相应的 Pod 容器；如果监听到修改 Pod 信息，则 kubelet 监听到变化后，会相应地修改本节点的 Pod 容器。

另外一个交互场景是 kube-controller-manager 进程与 API Server 的交互。kube-controller-manager 中的 Node Controller 模块通过 API Server 提供的 Watch 接口，实时监控 Node 的信息，并做相应处理。



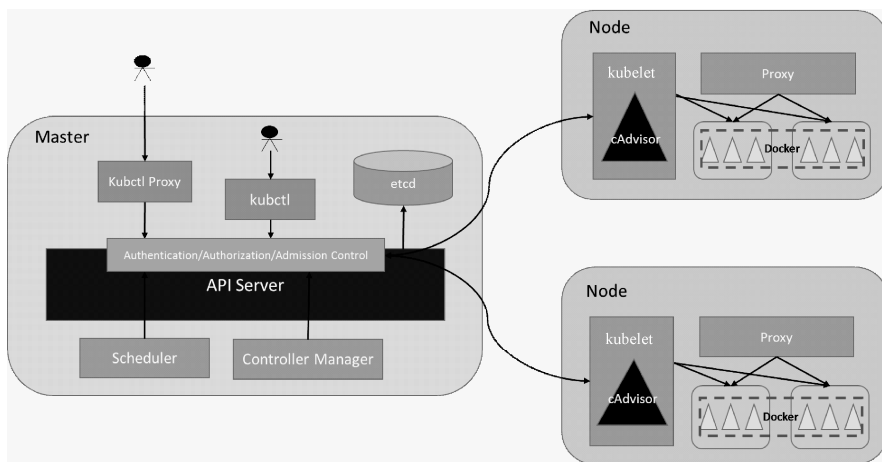


图 3.2 Kubernetes 结构图

还有一个比较重要的交互场景是 kube-scheduler 与 API Server 的交互。当 Scheduler 通过 API Server 的 Watch 接口监听到新建 Pod 副本的信息后，它会检索所有符合该 Pod 要求的 Node 列表，开始执行 Pod 调度逻辑，调度成功后将 Pod 绑定到目标节点上。

为了缓解集群各模块对 API Server 的访问压力，各功能模块都采用缓存机制来缓存数据。各功能模块定时从 API Server 获取指定的资源对象信息（通过 LIST 及 WATCH 方法），然后将这些信息保存到本地缓存，功能模块在某些情况下不直接访问 API Server，而是通过访问缓存数据来间接访问 API Server。

## 3.2 Controller Manager 原理分析

Controller Manager 作为集群内部的管理控制中心，负责集群内的 Node、Pod 副本、服务端点（Endpoint）、命名空间（Namespace）、服务账号（ServiceAccount）、资源定额（ResourceQuota）等的管理，当某个 Node 意外宕机时，Controller Manager 会及时发现此故障并执行自动化修复流程，确保集群始终处于预期的工作状态。

如图 3.3 所示，Controller Manager 内部包含 Replication Controller、Node Controller、ResourceQuota Controller、Namespace Controller、ServiceAccount Controller、Token Controller、Service Controller 及 Endpoint Controller 等多个 Controller，每种 Controller 都负责一种具体的控制流程，而 Controller Manager 正是这些 Controller 的核心管理者。

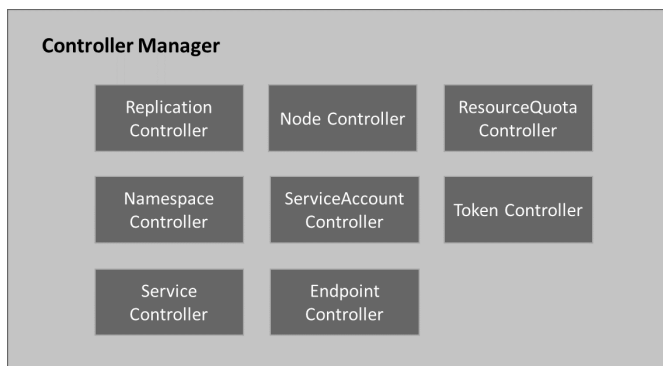


图 3.3 Controller Manager 结构图

一般来说，智能系统和自动系统通常会通过一个被称为“操纵系统”的机构来不断修正系统的工作状态。在 Kubernetes 集群中，每个 Controller 都是这样一个“操纵系统”，它们通过 API Server 提供的接口实时监控整个集群里的每个资源对象的当前状态，当发生各种故障导致系统状态发生变化时，会尝试着将系统状态从“现有状态”修正到“期望状态”。本节将详细分析 Controller Manager 的这些 Controller 的原理。

由于 ServiceAccount Controller 与 Token Controller 是与安全相关的两个控制器，并且与 Service Account、Token 密切相关，所以我们将对它们的分析放到后面的深入集群安全的章节中讲解。

在 Kubernetes 集群中与 Controller Manager 并重的另一个组件是 Kubernetes Scheduler，它的作用是将待调度的 Pod（包括通过 API Server 新创建的 Pod 及 RC 为补足副本而创建的 Pod 等）通过一些复杂的调度流程计算出最佳目标节点，然后绑定到该节点上。本章最后会介绍 Kubernetes Scheduler 调度器的基本原理。

### 3.2.1 Replication Controller

为了区分 Controller Manager 中的 Replication Controller(副本控制器)和资源对象 Replication Controller，我们将资源对象 Replication Controller 简写为 RC，而本节中的 Replication Controller 是指“副本控制器”，以便于后续分析。

Replication Controller 的核心作用是确保在任何时候集群中一个 RC 所关联的 Pod 副本数量保持预设值。如果发现 Pod 副本数量超过预期值，则 Replication Controller 会销毁一些 Pod 副本；反之，Replication Controller 会自动创建新的 Pod 副本，直到符合条件的 Pod 副本数量达到预设值。需要注意的一点是：只有当 Pod 的重启策略是 Always 时（RestartPolicy=Always），Replication

Controller 才会管理该 Pod 的操作（例如创建、销毁、重启等）。在通常情况下，Pod 对象被成功创建后不会消失，唯一的例外是当 Pod 处于 `succeeded` 或 `failed` 状态的时间过长（超时参数由系统设定）时，该 Pod 会被系统自动回收，管理该 Pod 的副本控制器将在其他工作节点上重新创建、运行该 Pod 副本。

RC 中的 Pod 模板就像一个模具，模具制作出来的东西一旦离开模具，它们之间就再也没关系了。同样，一旦 Pod 被创建完毕，无论模板如何变化，甚至换成一个新的模板，也不会影响到已经创建的 Pod。此外，Pod 可以通过修改它的标签来实现脱离 RC 的管控。该方法可以用于将 Pod 从集群中迁移、数据修复等调试。对于被迁移的 Pod 副本，RC 会自动创建一个新的副本替换被迁移的副本。需要注意的是，删除一个 RC 不会影响它所创建的 Pod。如果想删除一个 RC 所控制的 Pod，则需要将该 RC 的副本数（Replicas）属性设置为 0，这样所有的 Pod 副本都会被自动删除。

最好不要越过 RC 而直接创建 Pod，因为 Replication Controller 会通过 RC 管理 Pod 副本，实现自动创建、补足、替换、删除 Pod 副本，这样能提高系统的容灾能力，减少由于节点崩溃等意外状况造成的损失。即使你的应用程序只用到一个 Pod 副本，我们也强烈建议使用 RC 来定义 Pod。

总结一下 Replication Controller 的职责，如下所述。

- （1）确保当前集群中有且仅有  $N$  个 Pod 实例， $N$  是 RC 中定义的 Pod 副本数量。
- （2）通过调整 RC 的 `spec.replicas` 属性值来实现系统扩容或者缩容。
- （3）通过改变 RC 中的 Pod 模板（主要是镜像版本）来实现系统的滚动升级。

最后，我们总结一下 Replication Controller 的典型使用场景，如下所述。

（1）重新调度（Rescheduling）。如前面所提及的，不管你想运行 1 个副本还是 1000 个副本，副本控制器都能确保指定数量的副本存在于集群中，即使发生节点故障或 Pod 副本被终止运行等意外状况。

（2）弹性伸缩（Scaling）。手动或者通过自动扩容代理修改副本控制器的 `spec.replicas` 属性值，非常容易实现扩大或缩小副本的数量。

（3）滚动更新（Rolling Updates）。副本控制器被设计成通过逐个替换 Pod 的方式来辅助服务的滚动更新。推荐的方式是创建一个新的只有一个副本的 RC，若新的 RC 副本数量加 1，则旧的 RC 的副本数量减 1，直到这个旧的 RC 的副本数量为零，然后删除该旧的 RC。通过上述模式，即使在滚动更新的过程中发生了不可预料的错误，Pod 集合的更新也都在可控范围内。在理想情况下，滚动更新控制器需要将准备就绪的应用考虑在内，并保证在集群中任何时刻都有足够数量的可用 Pod。

### 3.2.2 Node Controller

kubelet 进程在启动时通过 API Server 注册自身的节点信息，并定时向 API Server 汇报状态信息，API Server 接收到这些信息后，将这些信息更新到 etcd 中，etcd 中存储的节点信息包括节点健康状况、节点资源、节点名称、节点地址信息、操作系统版本、Docker 版本、kubelet 版本等。节点健康状况包含“就绪”（True）“未就绪”（False）和“未知”（Unknown）三种。

Node Controller 通过 API Server 实时获取 Node 的相关信息，实现管理和监控集群中的各个 Node 节点的相关控制功能，Node Controller 的核心工作流程如图 3.4 所示。

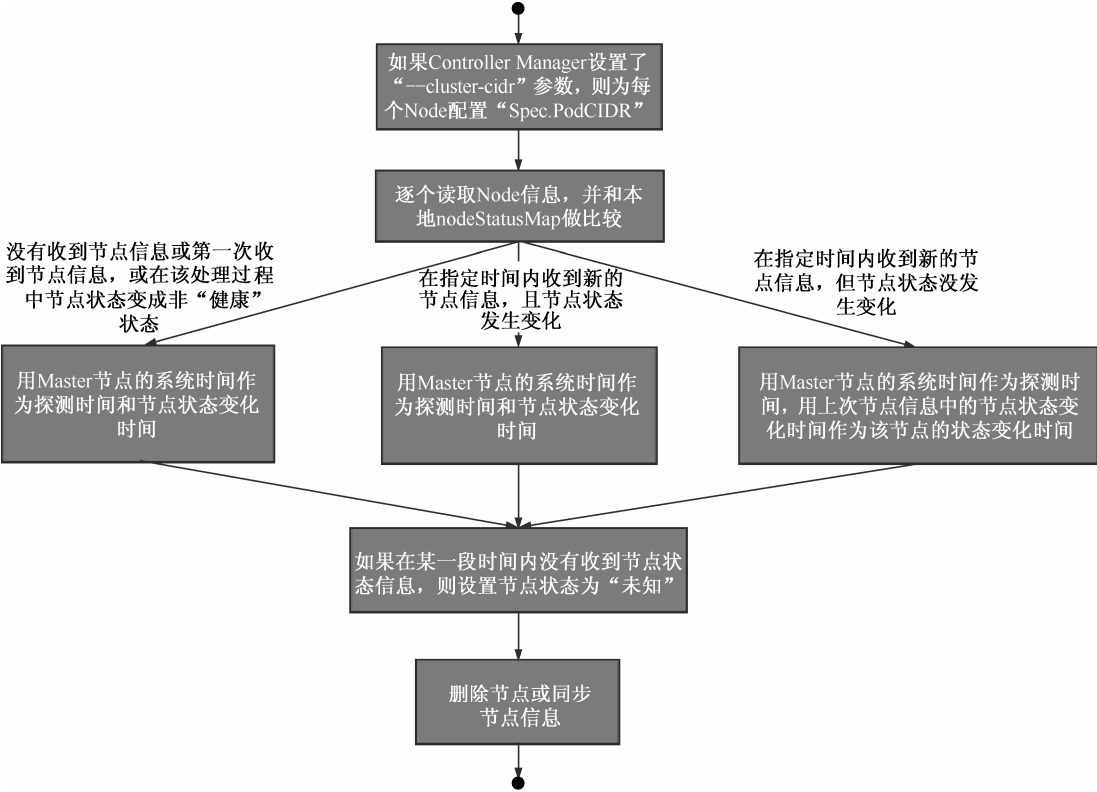


图 3.4 Node Controller 的核心工作流程

对流程中关键点的解释如下。

(1) Controller Manager 在启动时如果设置了 `--cluster-cidr` 参数，那么为每个没有设置 `Spec.PodCIDR` 的 Node 节点生成一个 CIDR 地址，并用该 CIDR 地址设置节点的 `Spec.PodCIDR` 属性，这样做的目的是防止不同节点的 CIDR 地址发生冲突。

(2) 逐个读取节点信息，多次尝试修改 `nodeStatusMap` 中的节点状态信息，将该节点信息和 `Node Controller` 的 `nodeStatusMap` 中保存的节点信息做比较。如果判断出没有收到 `kubelet` 发送的节点信息、第 1 次收到节点 `kubelet` 发送的节点信息，或在该处理过程中节点状态变成非“健康”状态，则在 `nodeStatusMap` 中保存该节点的状态信息，并用 `Node Controller` 所在节点的系统时间作为探测时间和节点状态变化时间。如果判断出在指定时间内收到新的节点信息，且节点状态发生变化，则在 `nodeStatusMap` 中保存该节点的状态信息，并用 `Node Controller` 所在节点的系统时间作为探测时间和节点状态变化时间。如果判断出在指定时间内收到新的节点信息，但节点状态没发生变化，则在 `nodeStatusMap` 中保存该节点的状态信息，并用 `Node Controller` 所在节点的系统时间作为探测时间，用上次节点信息中的节点状态变化时间作为该节点的状态变化时间。如果判断出在某一段时间（`gracePeriod`）内没有收到节点状态信息，则设置节点状态为“未知”（`Unknown`），并且通过 `API Server` 保存节点状态。

(3) 逐个读取节点信息，如果节点状态变为非“就绪”状态，则将节点加入待删除队列，否则将节点从该队列中删除。如果节点状态为非“就绪”状态，且系统指定了 `Cloud Provider`，则 `Node Controller` 调用 `Cloud Provider` 查看节点，若发现节点故障，则删除 `etcd` 中的节点信息，并删除和该节点相关的 `Pod` 等资源的信息。

### 3.2.3 ResourceQuota Controller

作为完备的企业级的容器集群管理平台，`Kubernetes` 也提供了资源配额管理（`ResourceQuota Controller`）这一高级功能，资源配额管理确保了指定的资源对象在任何时候都不会超量占用系统物理资源，避免了由于某些业务进程的设计或实现的缺陷导致整个系统运行紊乱甚至意外宕机，对整个集群的平稳运行和稳定性有非常重要的作用。

目前 `Kubernetes` 支持如下三个层次的资源配额管理。

- (1) 容器级别，可以对 `CPU` 和 `Memory` 进行限制。
- (2) `Pod` 级别，可以对一个 `Pod` 内所有容器的可用资源进行限制。
- (3) `Namespace` 级别，为 `Namespace`（多租户）级别的资源限制，包括：
  - ◎ `Pod` 数量；
  - ◎ `Replication Controller` 数量；
  - ◎ `Service` 数量；
  - ◎ `ResourceQuota` 数量；

- ◎ Secret 数量；
- ◎ 可持有的 PV（Persistent Volume）数量。

Kubernetes 的配额管理是通过 Admission Control（准入控制）来控制的，Admission Control 当前提供了两种方式的配额约束，分别是 LimitRanger 与 ResourceQuota。其中 LimitRanger 作用于 Pod 和 Container 上，而 ResourceQuota 则作用于 Namespace 上，限定一个 Namespace 里的各类资源的使用总额。

如图 3.5 所示，如果在 Pod 定义中同时声明了 LimitRanger，则用户通过 API Server 请求创建或修改资源时，Admission Control 会计算当前配额的使用情况，如果不符合配额约束，则创建对象失败。对于定义了 ResourceQuota 的 Namespace，ResourceQuota Controller 组件则负责定期统计和生成该 Namespace 下的各类对象的资源使用总量，统计结果包括 Pod、Service、RC、Secret 和 Persistent Volume 等对象实例个数，以及该 Namespace 下所有 Container 实例所使用的资源量（目前包括 CPU 和内存），然后将这些统计结果写入 etcd 的 resourceQuotaStatusStorage 目录（resourceQuotas/status）中。写入 resourceQuotaStatusStorage 的内容包含 Resource 名称、配额值（ResourceQuota 对象中 spec.hard 域下包含的资源值）、当前使用值（ResourceQuota Controller 统计出来的值）。随后这些统计信息被 Admission Control 使用，以确保相关 Namespace 下的资源配额总量不会超过 ResourceQuota 中的限定值。

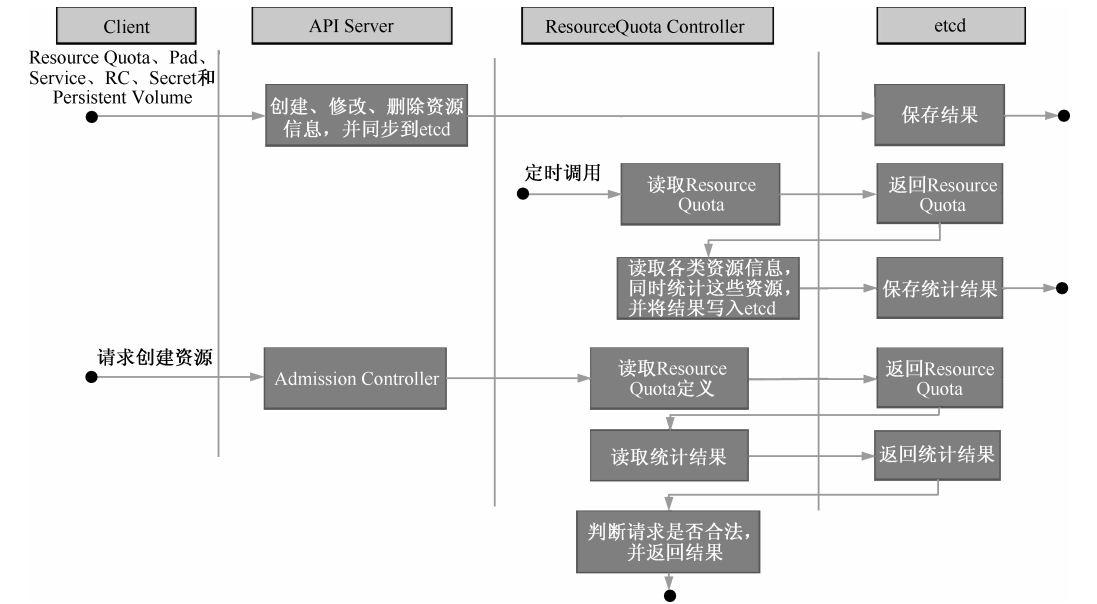


图 3.5 ResourceQuota Controller 流程图

### 3.2.4 Namespace Controller

用户通过 API Server 可以创建新的 Namespace 并保存在 etcd 中，Namespace Controller 定时通过 API Server 读取这些 Namespace 信息。如果 Namespace 被 API 标识为优雅删除（通过设置删除期限，即 DeletionTimestamp 属性被设置），则将该 Namespace 的状态设置成 “Terminating” 并保存到 etcd 中。同时 Namespace Controller 删除该 Namespace 下的 ServiceAccount、RC、Pod、Secret、PersistentVolume、ListRange、ResourceQuota 和 Event 等资源对象。

当 Namespace 的状态被设置成 “Terminating” 后，由 Admission Controller 的 NamespaceLifecycle 插件来阻止为该 Namespace 创建新的资源。同时，在 Namespace Controller 删除完该 Namespace 中的所有资源对象后，Namespace Controller 对该 Namespace 执行 finalize 操作，删除 Namespace 的 spec.finalizers 域中的信息。

如果 Namespace Controller 观察到 Namespace 设置了删除期限，同时 Namespace 的 spec.finalizers 域值是空的，那么 Namespace Controller 将通过 API Server 删除该 Namespace 资源。

### 3.2.5 Service Controller 与 Endpoint Controller

我们先说说 Endpoints Controller，在这之前，让我们先看看 Service、Endpoints 与 Pod 的关系，如图 3.6 所示，Endpoints 表示一个 Service 对应的所有 Pod 副本的访问地址，而 Endpoints Controller 就是负责生成和维护所有 Endpoints 对象的控制器。

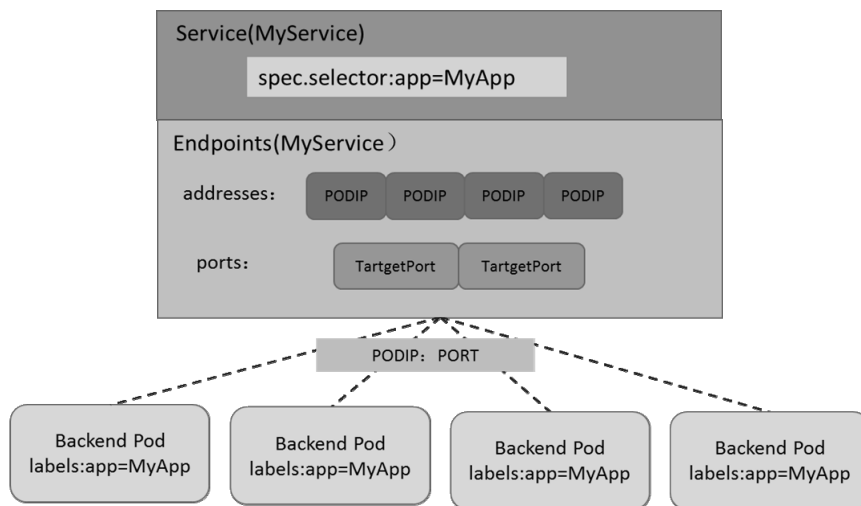


图 3.6 Service、Endpoint 与 Pod 的关系

它负责监听 Service 和对应的 Pod 副本的变化，如果监测到 Service 被删除，则删除和该 Service 同名的 Endpoints 对象。如果监测到新的 Service 被创建或者修改，则根据该 Service 信息获得相关的 Pod 列表，然后创建或者更新 Service 对应的 Endpoints 对象。如果监测到 Pod 的事件，则更新它所对应的 Service 的 Endpoints 对象（增加、删除或者修改对应的 Endpoint 条目）。

那么，Endpoints 对象是在哪里被使用的呢？答案是每个 Node 上的 kube-proxy 进程，kube-proxy 进程获取每个 Service 的 Endpoints，实现了 Service 的负载均衡功能。在后面的章节中我们会深入讲解这部分内容。

接下来我们说说 Service Controller 的作用，它其实是属于 Kubernetes 集群与外部的云平台之间的一个接口控制器。Service Controller 监听 Service 的变化，如果是一个 LoadBalancer 类型的 Service（externalLoadBalancers=true），则 Service Controller 确保外部的云平台上该 Service 对应的 LoadBalancer 实例被相应地创建、删除及更新路由转发表（根据 Endpoints 的条目）。

### 3.3 Scheduler 原理分析

我们在前面深入分析了 Controller Manager 及它所包含的各个组件的运行机制。本节我们将继续对 Kubernetes 中负责 Pod 调度的重要功能模块——Kubernetes Scheduler 的工作原理和运行机制做深入分析。

Kubernetes Scheduler 在整个系统中承担了“承上启下”的重要功能，“承上”是指它负责接收 Controller Manager 创建的新 Pod，为其安排一个落脚的“家”——目标 Node；“启下”是指安置工作完成后，目标 Node 上的 kubelet 服务进程接管后继工作，负责 Pod 生命周期中的“下半生”。

具体来说，Kubernetes Scheduler 的作用是将待调度的 Pod（API 新创建的 Pod、Controller Manager 为补足副本而创建的 Pod 等）按照特定的调度算法和调度策略绑定（Binding）到集群中的某个合适的 Node 上，并将绑定信息写入 etcd 中。在整个调度过程中涉及三个对象，分别是：待调度 Pod 列表、可用 Node 列表，以及调度算法和策略。简单地说，就是通过调度算法调度为待调度 Pod 列表的每个 Pod 从 Node 列表中选择一个最适合的 Node。

随后，目标节点上的 kubelet 通过 API Server 监听到 Kubernetes Scheduler 产生的 Pod 绑定事件，然后获取对应的 Pod 清单，下载 Image 镜像，并启动容器。完整的流程如图 3.7 所示。



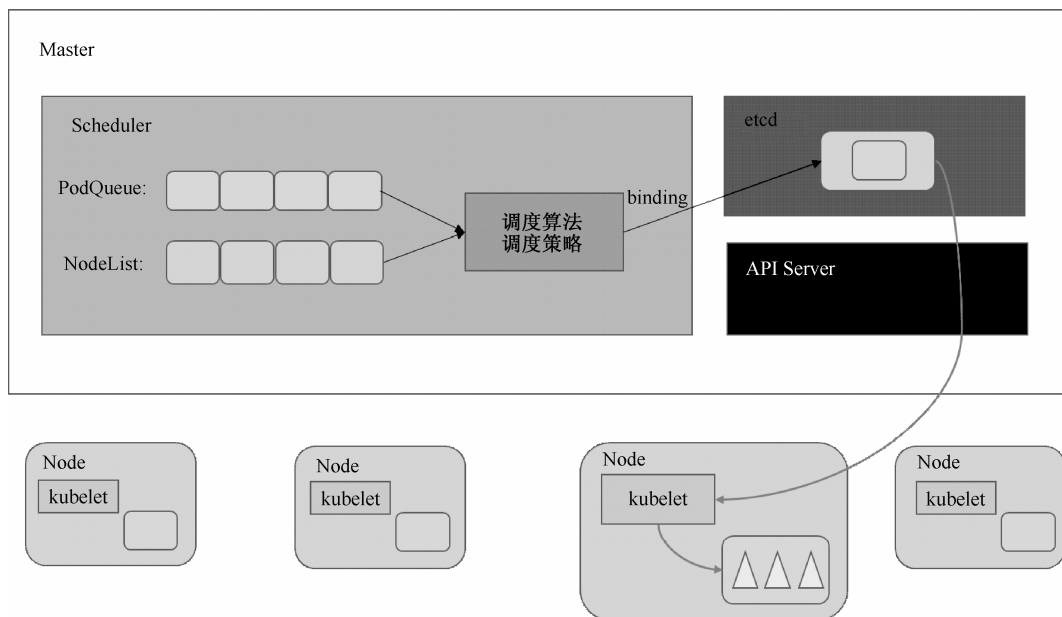


图 3.7 Scheduler 流程

Kubernetes Scheduler 当前提供的默认调度流程分为以下两步。

(1) 预选调度过程，即遍历所有目标 Node，筛选出符合要求的候选节点。为此，Kubernetes 内置了多种预选策略（xxx Predicates）供用户选择。

(2) 确定最优节点，在第 1 步的基础上，采用优选策略（xxx Priority）计算出每个候选节点的积分，积分最高者胜出。

Kubernetes Scheduler 的调度流程是通过插件方式加载的“调度算法提供者”（AlgorithmProvider）具体实现的。一个 AlgorithmProvider 其实就是包括了一组预选策略与一组优先选择策略的结构体，注册 AlgorithmProvider 的函数如下：

```
func RegisterAlgorithmProvider(name string, predicateKeys, priorityKeys
util.StringSet)
```

它包含三个参数：“name string”参数为算法名；“predicateKeys”参数为算法用到的预选策略集合；“priorityKeys”为算法用到的优选策略集合。

Scheduler 中可用的预选策略包含：NoDiskConflict、PodFitsResources、PodSelectorMatches、PodFitsHost、CheckNodeLabelPresence、CheckServiceAffinity 和 PodFitsPorts 策略等。其默认的 AlgorithmProvider 加载的预选策略 Predicates 包括：PodFitsPorts（PodFitsPorts）、PodFitsResources（PodFitsResources）、NoDiskConflict（NoDiskConflict）、MatchNodeSelector（PodSelectorMatches）

和 `HostName` (`PodFitsHost`)，即每个节点只有通过前面提及的 5 个默认预选策略后，才能初步被选中，进入下一个流程。

下面列出的是对所有预选策略的详细说明。

### 1) NoDiskConflict

判断备选 Pod 的 `gcePersistentDisk` 或 `AWSElasticBlockStore` 和备选的节点中已存在的 Pod 是否存在冲突。检测过程如下。

(1) 首先，读取备选 Pod 的所有 Volume 的信息（即 `pod.Spec.Volumes`），对每个 Volume 执行以下步骤进行冲突检测。

(2) 如果该 Volume 是 `gcePersistentDisk`，则将 Volume 和备选节点上的所有 Pod 的每个 Volume 进行比较，如果发现相同的 `gcePersistentDisk`，则返回 `false`，表明存在磁盘冲突，检查结束，反馈给调度器该备选节点不适合作为备选 Pod；如果该 Volume 是 `AWSElasticBlockStore`，则将 Volume 和备选节点上的所有 Pod 的每个 Volume 进行比较，如果发现相同的 `AWSElasticBlockStore`，则返回 `false`，表明存在磁盘冲突，检查结束，反馈给调度器该备选节点不适合备选 Pod。

(3) 如果检查完备选 Pod 的所有 Volume 均未发现冲突，则返回 `true`，表明不存在磁盘冲突，反馈给调度器该备选节点适合备选 Pod。

### 2) PodFitsResources

判断备选节点的资源是否满足备选 Pod 的需求，检测过程如下。

(1) 计算备选 Pod 和节点中已存在 Pod 的所有容器的需求资源（内存和 CPU）的总和。

(2) 获得备选节点的状态信息，其中包含节点的资源信息。

(3) 如果备选 Pod 和节点中已存在 Pod 的所有容器的需求资源（内存和 CPU）的总和，超出了备选节点拥有的资源，则返回 `false`，表明备选节点不适合备选 Pod，否则返回 `true`，表明备选节点适合备选 Pod。

### 3) PodSelectorMatches

判断备选节点是否包含备选 Pod 的标签选择器指定的标签。

(1) 如果 Pod 没有指定 `spec.nodeSelector` 标签选择器，则返回 `true`。

(2) 否则，获得备选节点的标签信息，判断节点是否包含备选 Pod 的标签选择器（`spec.nodeSelector`）所指定的标签，如果包含，则返回 `true`，否则返回 `false`。

### 4) PodFitsHost

判断备选 Pod 的 `spec.nodeName` 域所指定的节点名称和备选节点的名称是否一致，如果一致，则返回 `true`，否则返回 `false`。

### 5) CheckNodeLabelPresence

如果用户在配置文件中指定了该策略，则 Scheduler 会通过 RegisterCustomFitPredicate 方法注册该策略。该策略用于判断策略列出的标签在备选节点中存在时，是否选择该备选节点。

(1) 读取备选节点的标签列表信息。

(2) 如果策略配置的标签列表存在于备选节点的标签列表中，且策略配置的 presence 值为 false，则返回 false，否则返回 true；如果策略配置的标签列表不存在于备选节点的标签列表中，且策略配置的 presence 值为 true，则返回 false，否则返回 true。

### 6) CheckServiceAffinity

如果用户在配置文件中指定了该策略，则 Scheduler 会通过 RegisterCustomFitPredicate 方法注册该策略。该策略用于判断备选节点是否包含策略指定的标签，或包含和备选 Pod 在相同 Service 和 Namespace 下的 Pod 所在节点的标签列表。如果存在，则返回 true，否则返回 false。

### 7) PodFitsPorts

判断备选 Pod 所用的端口列表中的端口是否在备选节点中已被占用，如果被占用，则返回 false，否则返回 true。

Scheduler 中的优选策略包含：LeastRequestedPriority、CalculateNodeLabelPriority 和 BalancedResourceAllocation 等。每个节点通过优先选择策略时都会算出一个得分，计算各项得分，最终选出得分值最大的节点作为优选的结果（也是调度算法的结果）。

下面是对所有优选策略的详细说明。

#### 1) LeastRequestedPriority

该优选策略用于从备选节点列表中选出资源消耗最小的节点。

(1) 计算出所有备选节点上运行的 Pod 和备选 Pod 的 CPU 占用量 totalMilliCPU。

(2) 计算出所有备选节点上运行的 Pod 和备选 Pod 的内存占用量 totalMemory。

(3) 计算每个节点的得分，计算规则大致如下。

NodeCpuCapacity 为节点 CPU 计算能力，NodeMemoryCapacity 为节点内存大小。

```
score=int(((nodeCpuCapacity-totalMilliCPU)*10)/ nodeCpuCapacity+((nodeMemoryCapacity-totalMemory)*10)/ nodeCpuMemory)/2)
```

#### 2) CalculateNodeLabelPriority

如果用户在配置文件中指定了该策略，则 scheduler 会通过 RegisterCustomPriorityFunction 方法注册该策略。该策略用于判断策略列出的标签在备选节点中存在时，是否选择该备选节点。如果备选节点的标签在优选策略的标签列表中且优选策略的 presence 值为 true，或者备选节点

的标签不在优选策略的标签列表中且优选策略的 `presence` 值为 `false`，则备选节点 `score=10`，否则备选节点 `score=0`。

### 3) **BalancedResourceAllocation**

该优选策略用于从备选节点列表中选出各项资源使用率最均衡的节点。

(1) 计算出所有备选节点上运行的 Pod 和备选 Pod 的 CPU 占用量 `totalMilliCPU`。

(2) 计算出所有备选节点上运行的 Pod 和备选 Pod 的内存占用量 `totalMemory`。

(3) 计算每个节点的得分，计算规则大致如下。

`NodeCpuCapacity` 为节点 CPU 计算能力，`NodeMemoryCapacity` 为节点内存大小。

```
score= int(10-math.Abs(totalMilliCPU/nodeCpuCapacity-totalMemory/  
nodeMemoryCapacity)*10)
```

## 3.4 kubelet 运行机制分析

在 Kubernetes 集群中，在每个 Node 节点（又称 Minion）上都会启动一个 kubelet 服务进程。该进程用于处理 Master 节点下发到本节点的任务，管理 Pod 及 Pod 中的容器。每个 kubelet 进程会在 API Server 上注册节点自身信息，定期向 Master 节点汇报节点资源的使用情况，并通过 cAdvisor 监控容器和节点资源。

### 3.4.1 节点管理

节点通过设置 kubelet 的启动参数 “`--register-node`”，来决定是否向 API Server 注册自己。如果该参数的值为 `true`，那么 kubelet 将试着通过 API Server 注册自己。在自注册时，kubelet 启动时还包含下列参数。

- ◎ `--api-servers`: API Server 的位置。
- ◎ `--kubeconfig`: kubeconfig 文件，用于访问 API Server 的安全配置文件。
- ◎ `--cloud-provider`: 云服务商（IaaS）地址，仅用于公有云环境。

当前每个 kubelet 被授予创建和修改任何节点的权限。但是在实践中，它仅仅创建和修改自己。将来，我们计划限制 kubelet 的权限，仅允许它修改和创建其所在节点的权限。如果在集群运行过程中遇到集群资源不足的情况，则用户很容易通过添加机器及运用 kubelet 的自注册模式来实现扩容。

在某些情况下，Kubernetes 集群中的某些 kubelet 没有选择自注册模式，用户需要自己去配置 Node 的资源信息，同时告知 Node 上的 kubelet API Server 的位置。集群管理者能够创建和修改节点信息。如果管理者希望手动创建节点信息，则通过设置 kubelet 的启动参数 “--register-node=false” 即可。

kubelet 在启动时通过 API Server 注册节点信息，并定时向 API Server 发送节点的新消息，API Server 在接收到这些信息后，将这些信息写入 etcd。通过 kubelet 的启动参数 “--node-status-update-frequency” 设置 kubelet 每隔多长时间向 API Server 报告节点状态，默认为 10s。

### 3.4.2 Pod 管理

kubelet 通过以下几种方式获取自身 Node 上所要运行的 Pod 清单。

(1) 文件：kubelet 启动参数 “--config” 指定的配置文件目录下的文件（默认目录为 “/etc/kubernetes/manifests/”）。通过 --file-check-frequency 设置检查该文件目录的时间间隔，默认为 20s。

(2) HTTP 端点 (URL)：通过 “--manifest-url” 参数设置。通过 --http-check-frequency 设置检查该 HTTP 端点数据的时间间隔，默认为 20s。

(3) API Server：kubelet 通过 API Server 监听 etcd 目录，同步 Pod 列表。

所有以非 API Server 方式创建的 Pod 都叫作 Static Pod。kubelet 将 Static Pod 的状态汇报给 API Server，API Server 为该 Static Pod 创建一个 Mirror Pod 和其相匹配。Mirror Pod 的状态将真实反映 Static Pod 的状态。当 Static Pod 被删除时，与之相对应的 Mirror Pod 也会被删除。在本章中我们只讨论通过 API Server 获得 Pod 清单的方式。kubelet 通过 API Server Client 使用 Watch 加 List 的方式监听 “/registry/nodes/\$当前节点的名称” 和 “/registry/pods” 目录，将获取的信息同步到本地缓存中。

kubelet 监听 etcd，所有针对 Pod 的操作将会被 kubelet 监听到。如果发现新的绑定到本节点的 Pod，则按照 Pod 清单的要求创建该 Pod。

如果发现本地的 Pod 被修改，则 kubelet 会做出相应的修改，比如删除 Pod 中的某个容器时，则通过 Docker Client 删除该容器。

如果发现删除本节点的 Pod，则删除相应的 Pod，并通过 Docker Client 删除 Pod 中的容器。

kubelet 读取监听到的信息，如果是创建和修改 Pod 任务，则做如下处理。

(1) 为该 Pod 创建一个数据目录。

(2) 从 API Server 读取该 Pod 清单。

(3) 为该 Pod 挂载外部卷（External Volume）。

(4) 下载 Pod 用到的 Secret。

(5) 检查已经运行在节点中的 Pod，如果该 Pod 没有容器或 Pause 容器（“kubernetes/pause”镜像创建的容器）没有启动，则先停止 Pod 里所有容器的进程。如果在 Pod 中有需要删除的容器，则删除这些容器。

(6) 用“kubernetes/pause”镜像为每个 Pod 创建一个容器。该 Pause 容器用于接管 Pod 中所有其他容器的网络。每创建一个新的 Pod，kubelet 都会先创建一个 Pause 容器，然后创建其他容器。“kubernetes/pause”镜像大概为 200KB，是一个非常小的容器镜像。

(7) 为 Pod 中的每个容器做如下处理。

- ◎ 为容器计算一个 hash 值，然后用容器的名字去查询对应 Docker 容器的 hash 值。若查找到容器，且两者的 hash 值不同，则停止 Docker 中容器的进程，并停止与之关联的 Pause 容器的进程；若两者相同，则不做任何处理。
- ◎ 如果容器被终止了，且容器没有指定的 restartPolicy（重启策略），则不做任何处理。
- ◎ 调用 Docker Client 下载容器镜像，调用 Docker Client 运行容器。

### 3.4.3 容器健康检查

Pod 通过两类探针来检查容器的健康状态。一个是 LivenessProbe 探针，用于判断容器是否健康，告诉 kubelet 一个容器什么时候处于不健康的状态。如果 LivenessProbe 探针探测到容器不健康，则 kubelet 将删除该容器，并根据容器的重启策略做相应的处理。如果一个容器不包含 LivenessProbe 探针，那么 kubelet 认为该容器的 LivenessProbe 探针返回的值永远是“Success”；另一类是 ReadinessProbe 探针，用于判断容器是否启动完成，且准备接收请求。如果 ReadinessProbe 探针检测到失败，则 Pod 的状态将被修改，Endpoint Controller 将从 Service 的 Endpoint 中删除包含该容器所在 Pod 的 IP 地址的 Endpoint 条目。

kubelet 定期调用容器中的 LivenessProbe 探针来诊断容器的健康状况。LivenessProbe 包含以下三种实现方式。

(1) ExecAction：在容器内部执行一个命令，如果该命令的退出状态码为 0，则表明容器健康。

(2) TCPSocketAction：通过容器的 IP 地址和端口号执行 TCP 检查，如果端口能被访问，则表明容器健康。

(3) HTTPGetAction：通过容器的 IP 地址和端口号及路径调用 HTTP Get 方法，如果响应

的状态码大于等于 200 且小于等于 400，则认为容器状态健康。

LivenessProbe 探针包含在 Pod 定义的 `spec.containers.{某个容器}` 中。下面的例子展示了两种 Pod 中容器健康检查的方式：HTTP 检查和容器命令执行检查。下面所列的内容实现了通过容器命令执行检查：

```
livenessProbe:
  exec:
    command:
      - cat
      - /tmp/health
  initialDelaySeconds: 15
  timeoutSeconds: 1
```

kubelet 在容器中执行 “cat /tmp/health” 命令，如果该命令返回的值为 0，则表明容器处于健康状态，否则表明容器处于不健康状态。

下面所列的内容实现了容器的 HTTP 检查：

```
livenessProbe:
  httpGet:
    path: /healthz
    port: 8080
  initialDelaySeconds: 15
  timeoutSeconds: 1
```

kubelet 发送一个 HTTP 请求到本地主机和端口及指定的路径，来检查容器的健康状况。

### 3.4.4 cAdvisor 资源监控

在 Kubernetes 集群中如何监控资源的使用情况？

在 Kubernetes 集群中，应用程序的执行情况可以在不同的级别上监测到，这些级别包括：容器、Pod、Service 和整个集群。作为 Kubernetes 集群的一部分，Kubernetes 希望提供给用户详细的各个级别的资源使用信息，这将使用户能够深入地了解应用的执行情况，并找到应用中可能的瓶颈。Heapster 项目为 Kubernetes 提供了一个基本的监控平台，它是集群级别的监控和事件数据集成器(Aggregator)。Heapster 作为 Pod 运行在 Kubernetes 集群中，和运行在 Kubernetes 集群中的其他应用相似。Heapster Pod 通过 kubelet（运行在节点上的 Kubernetes 代理）发现所有运行在集群中的节点，并查看来自这些节点的资源使用状况信息。kubelet 通过 cAdvisor 获取其所在节点及容器的数据，Heapster 通过带着关联标签的 Pod 分组这些信息，这些数据被推到一个可配置的后端，用于存储和可视化展示。当前支持的后端包括 InfluxDB（with Grafana for Visualization）和 Google Cloud Monitoring。

cAdvisor 是一个开源的分析容器资源使用率和性能特性的代理工具。它是因为容器而产生的，因此自然支持 Docker 容器。在 Kubernetes 项目中，cAdvisor 被集成到 Kubernetes 代码中。cAdvisor 自动查找所有在其所在节点上的容器，自动采集 CPU、内存、文件系统和网络使用的统计信息。cAdvisor 通过它所在节点机的 Root 容器，采集并分析该节点机的全面使用情况。

在大部分 Kubernetes 集群中，cAdvisor 通过它所在节点机的 4194 端口暴露一个简单的 UI。如图 3.8 所示是 cAdvisor 的一个截图。

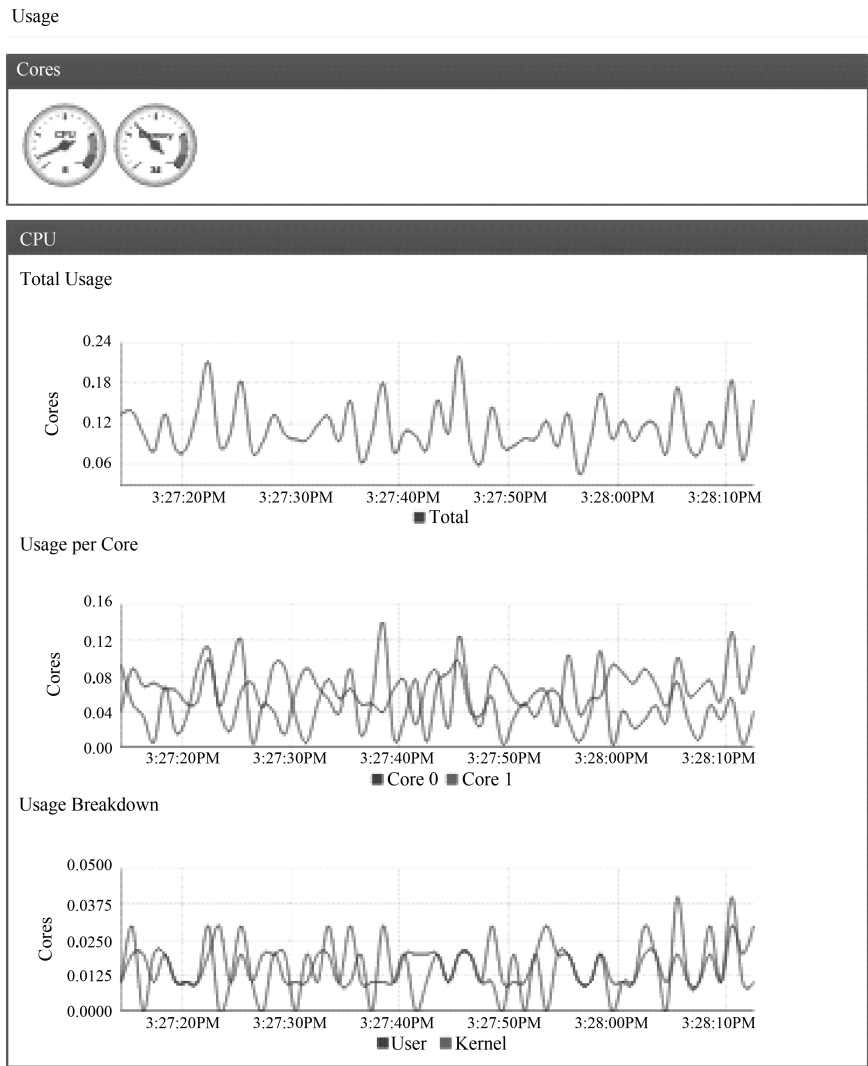


图 3.8 cAdvisor 的一个截图



kubelet 作为连接 Kubernetes Master 和各节点机之间的桥梁，管理运行在节点机上的 Pod 和容器。kubelet 将每个 Pod 转换成它的成员容器，同时从 cAdvisor 获取单独的容器使用统计信息，然后通过该 REST API 暴露这些聚合后的 Pod 资源使用的统计信息。

## 3.5 kube-proxy 运行机制分析

我们在前面已经了解到，为了支持集群的水平扩展、高可用性，Kubernetes 抽象出了 Service 的概念。Service 是对一组 Pod 的抽象，它会根据访问策略（如负载均衡策略）来访问这组 Pod。

Kubernetes 在创建服务时会为服务分配一个虚拟的 IP 地址，客户端通过访问这个虚拟的 IP 地址来访问服务，而服务则负责将请求转发到后端的 Pod 上。这不就是一个反向代理吗？不错，这就是一个反向代理。但是，它和普通的反向代理有一些不同：首先它的 IP 地址是虚拟的，想从外面访问还需要一些技巧；其次是它的部署和启停是 Kubernetes 统一自动管理的。

Service 在很多情况下只是一个概念，而真正将 Service 的作用落实的是背后的 kube-proxy 服务进程。只有理解了 kube-proxy 的原理和机制，我们才能真正理解 Service 背后的实现逻辑。

在 Kubernetes 集群的每个 Node 上都会运行一个 kube-proxy 服务进程，这个进程可以看作 Service 的透明代理兼负载均衡器，其核心功能是将到某个 Service 的访问请求转发到后端的多个 Pod 实例上。对每一个 TCP 类型的 Kubernetes Service，kube-proxy 都会在本地 Node 上建立一个 SocketServer 来负责接收请求，然后均匀发送到后端某个 Pod 的端口上，这个过程默认采用 Round Robin 负载均衡算法。另外，Kubernetes 也提供通过修改 Service 的 service.spec.sessionAffinity 参数的值来实现会话保持特性的定向转发，如果设置的值为“ClientIP”，则来自同一个 ClientIP 的请求都转发到同一个后端 Pod 上。

此外，Service 的 Cluster IP 与 NodePort 等概念是 kube-proxy 服务通过 Iptables 的 NAT 转换实现的，kube-proxy 在运行过程中动态创建与 Service 相关的 Iptables 规则，这些规则实现了 Cluster IP 及 NodePort 的请求流量重定向到 kube-proxy 进程上对应服务的代理端口的功能。由于 Iptables 机制针对的是本地的 kube-proxy 端口，所以每个 Node 上都要运行 kube-proxy 组件，这样一来，在 Kubernetes 集群内部，我们可以在任意 Node 上发起对 Service 的访问请求。

综上所述，由于 kube-proxy 的作用，在 Service 的调用过程中客户端无须关心后端有几个 Pod，中间过程的通信、负载均衡及故障恢复都是透明的，如图 3.9 所示。

访问 Service 的请求，不论是用 Cluster IP + TargetPort 的方式，还是用节点机 IP + NodePort 的方式，都被节点机的 Iptables 规则重定向到 kube-proxy 监听 Service 服务代理端口。kube-proxy 接收到 Service 的访问请求后，会如何选择后端的 Pod 呢？

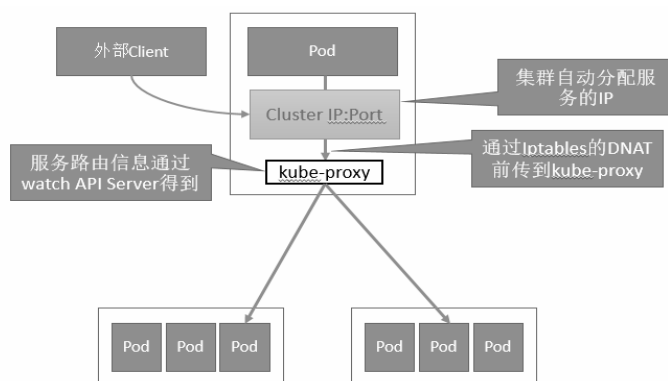


图 3.9 Service 的负载均衡转发规则

首先，目前 kube-proxy 的负载均衡器只支持 Round Robin 算法。Round Robin 算法按照成员列表逐个选取成员，如果一轮循环完，便从头开始下一轮，如此循环往复。kube-proxy 的负载均衡器在 Round Robin 算法的基础上还支持 Session 保持。如果 Service 在定义中指定了 Session 保持，则 kube-proxy 接收请求时会从本地内存中查找是否存在来自该请求 IP 的 affinityState 对象，如果存在该对象，且 Session 没有超时，则 kube-proxy 将请求转向该 affinityState 所指向的后端 Pod。如果本地存在没有来自该请求 IP 的 affinityState 对象，则按照 Round Robin 算法为该请求挑选一个 Endpoint，并创建一个 affinityState 对象，记录请求的 IP 和指向的 Endpoint。后面的请求就会“黏连”到这个创建好的 affinityState 对象上，这就实现了客户端 IP 会话保持的功能。

接下来我们深入分析 kube-proxy 的实现细节。

kube-proxy 通过查询和监听 API Server 中 Service 与 Endpoints 的变化，为每个 Service 都建立了一个“服务代理对象”，并自动同步。服务代理对象是 kube-proxy 程序内部的一种数据结构，它包括一个用于监听此服务请求的 SocketServer，SocketServer 的端口是随机选择的一个本地空闲端口。此外，kube-proxy 内部也创建了一个负载均衡器——LoadBalancer，LoadBalancer 上保存了 Service 到对应的后端 Endpoint 列表的动态转发路由表，而具体的路由选择则取决于 Round Robin 负载均衡算法及 Service 的 Session 会话保持（SessionAffinity）这两个特性。

针对发生变化的 Service 列表，kube-proxy 会逐个处理。下面是具体的处理流程。

（1）如果该 Service 没有设置集群 IP（ClusterIP），则不做任何处理，否则，获取该 Service 的所有端口定义列表（spec.ports 域）。

（2）逐个读取服务端口定义列表中的端口信息，根据端口名称、Service 名称和 Namespace 判断本地是否已经存在对应的服务代理对象，如果不存在则新建；如果存在并且 Service 端口被修改过，则先删除 iptables 中和该 Service 端口相关的规则，关闭服务代理对象，然后走新建流程，即为该 Service 端口分配服务代理对象并为该 Service 创建相关的 iptables 规则。

(3) 更新负载均衡器组件中对应 Service 的转发地址列表，对于新建的 Service，确定转发时的会话保持策略。

(4) 对于已经删除的 Service 则进行清理。

而针对 Endpoint 的变化，kube-proxy 会自动更新负载均衡器中对应 Service 的转发地址列表。

下面讲解 kube-proxy 针对 Iptables 所做的一些细节操作。

kube-proxy 在启动时和监听到 Service 或 Endpoint 的变化后，会在本机 Iptables 的 NAT 表中添加 4 条规则链。

(1) KUBE-PORTALS-CONTAINER：从容器中通过 Service Cluster IP 和端口号访问 Service 的请求。

(2) KUBE-PORTALS-HOST：从主机中通过 Service Cluster IP 和端口号访问 Service 的请求。

(3) KUBE-NODEPORT-CONTAINER：从容器中通过 Service 的 NodePort 端口号访问 Service 的请求。

(4) KUBE-NODEPORT-HOST：从主机中通过 Service 的 NodePort 端口号访问 Service 的请求。

此外，kube-proxy 在 Iptables 中为每个 Service 创建由 Cluster IP + Service 端口到 kube-proxy 所在主机 IP + Service 代理服务所监听的端口的转发规则。转发规则的包匹配规则部分 (CRETIRIA) 如下所示：

```
-m comment --comment $SERVICESTRING -p $PROTOCOL -m $PROTOCOL --dport $DESTPORT
-d $DESTIP
```

其中，“-m comment --comment”表示匹配规则使用 Iptables 的显式扩展的注释功能；“\$SERVICESTRING”为注释的内容；“-p \$PROTOCOL -m \$PROTOCOL --dport \$DESTPORT -d \$DESTIP”表示协议为“\$PROTOCOL”且目标地址和端口为“\$DESTIP”和“\$DESTPORT”的包，其中，“\$PROTOCOL”可以为 TCP 或 UDP，“\$DESTIP”和“\$DESTPORT”为 Service 的 Cluster IP 和 TargetPort。

对于转发规则的跳转部分 (-j 部分)，如果请求来自本地容器，且 Service 代理服务监听的是所有的接口（例如 IPv4 的地址为 0.0.0.0），则跳转部分如下所示：

```
-j REDIRECT --to-ports $proxyPort
```

其表示该规则的功能是实现数据包的端口重定向，重定向到 \$proxyPort 端口（Service 代理服务监听的端口）；否则，跳转部分如下所示：

```
-j DNAT --to-destination proxyIP:proxyPort
```

表示该规则的功能是实现数据包转发，数据包的目的地址变为“proxyIP:proxyPort”（即 Service 代理服务所在的 IP 地址和端口，这些地址和端口都会被替换成实际的地址和端口）。

如果 Service 类型为 NodePort，则 kube-proxy 在 Iptables 中除了添加上面提及的规则，还会为每个 Service 创建由 NodePort 端口到 kube-proxy 所在主机 IP + Service 代理服务所监听的端口的转发规则。转发规则的包匹配规则部分（CRETIRIA）如下所示：

```
-m comment --comment $SERVICESTRING -p $PROTOCOL -m $PROTOCOL --dport $NODEPORT
```

上面所列的内容用于匹配目的端口为“\$NODEPORT”的包。

转发规则的跳转部分（-j 部分）和前面提及的跳转规则一致。

最后，我们以本书第 2 章的 Hello World 为例，看看 kube-proxy 为 redis-master 服务所生成的 Iptables 转发规则：

```
$ iptables-save | grep redis-master
-A KUBE-PORTALS-CONTAINER -d 10.254.208.57/32 -p tcp -m comment --comment
"default/redis-master:" -m tcp --dport 6379 -j REDIRECT --to-ports 42872
-A KUBE-PORTALS-HOST -d 10.254.208.57/32 -p tcp -m comment --comment
"default/redis-master:" -m tcp --dport 6379 -j DNAT --to-destination
192.168.1.130:42872
```

可以看到，对“redis-master” Service 的 6379 端口的访问将会被转发到物理机的 42872 端口上。而 42872 端口就是 kube-proxy 为这个 Service 打开的随机本地端口。

最后，给出一个总结性的示意图，如图 3.10 所示。

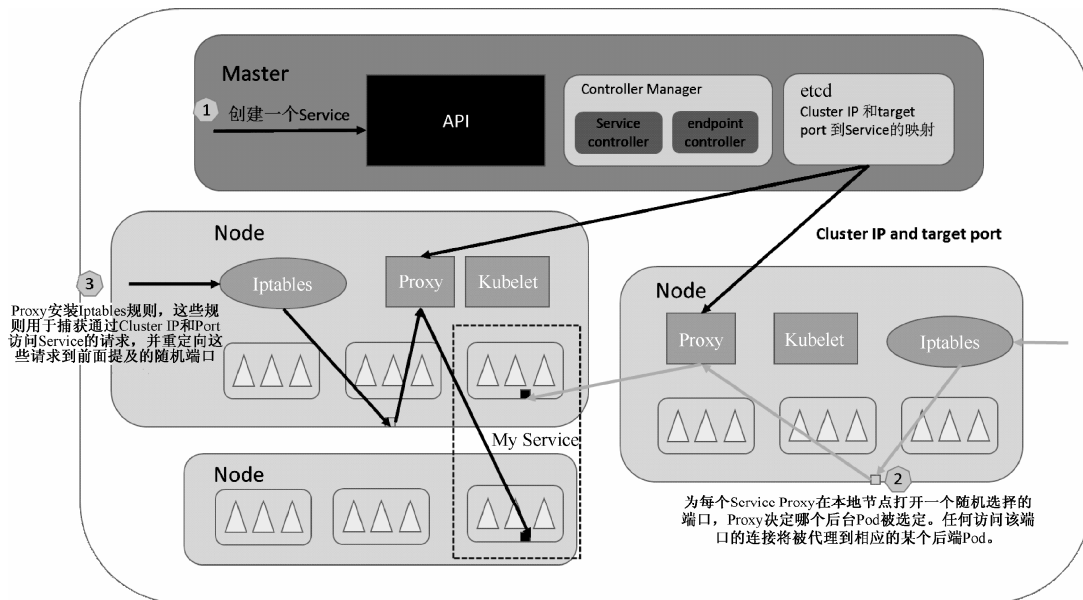


图 3.10 kube-proxy 工作原理示意图

## 3.6 深入分析集群安全机制

Kubernetes 通过一系列机制来实现集群的安全控制，其中包括 API Server 的认证授权、准入控制机制及保护敏感信息的 Secret 机制等。集群的安全性必须考虑如下几个目标。

- (1) 保证容器与其所在的宿主机的隔离。
- (2) 限制容器给基础设施及其他容器带来消极影响的能力。
- (3) 最小权限原则——合理限制所有组件的权限，确保组件只执行它被授权的行为，通过限制单个组件的能力来限制它所能到达的权限范围。
- (4) 明确组件间边界的划分。
- (5) 划分普通用户和管理员的角色。
- (6) 在必要时允许将管理员权限赋给普通用户。
- (7) 允许拥有“Secret”数据（Keys、Certs、Passwords）的应用在集群中运行。

下面分别从 Authentication、Authorization、Admission Control、Secret 和 Service Account 等方面来说明集群的安全机制。

### 3.6.1 API Server 认证管理（Authentication）

我们知道，Kubernetes 集群中所有资源的访问和变更都是通过 Kubernetes API Server 的 REST API 来实现的，所以集群安全的关键点就在于如何识别并认证客户端身份（Authentication），以及随后访问权限的授权（Authorization）这两个关键问题，本节对认证管理进行说明。

我们知道，Kubernetes 集群提供了 3 种级别的客户端身份认证方式。

- ◎ 最严格的 HTTPS 证书认证：基于 CA 根证书签名的双向数字证书认证方式。
- ◎ HTTP Token 认证：通过一个 Token 来识别合法用户。
- ◎ HTTP Base 认证：通过用户名+密码的方式认证。

首先，我们说说 HTTPS 证书认证的原理。

这里需要有一个 CA 证书，我们知道 CA 是 PKI 系统中通信双方都信任的实体，被称为可信第三方（Trusted Third Party, TTP）。CA 作为可信第三方的重要条件之一就是 CA 的行为具有非否认性。作为第三方而不是简单的上级，就必须能让信任者有追究自己责任的能力。CA 通过证书证实他人的公钥信息，证书上有 CA 的签名。用户如果因为信任证书而有了损失，则

证书可以作为有效的证据用于追究 CA 的法律责任。正是因为 CA 承担责任的承诺，所以 CA 也被称为可信第三方。在很多情况下，CA 与用户是相互独立的实体，CA 作为服务提供方，有可能因为服务质量问题（例如，发布的公钥数据有错误）而给用户带来损失。在证书中绑定了公钥数据和相应私钥拥有者的身份信息，并带有 CA 的数字签名；证书中也包含了 CA 的名称，以便于依赖方找到 CA 的公钥，验证证书上的数字签名。

CA 认证涉及诸多概念，比如根证书、自签名证书、密钥、私钥、加密算法及 HTTPS 等，本书大致讲述 SSL 协议的流程，有助于对 CA 认证和 Kubernetes CA 认证的配置过程的理解。

如图 3.11 所示，CA 认证大概包含下面几个步骤。

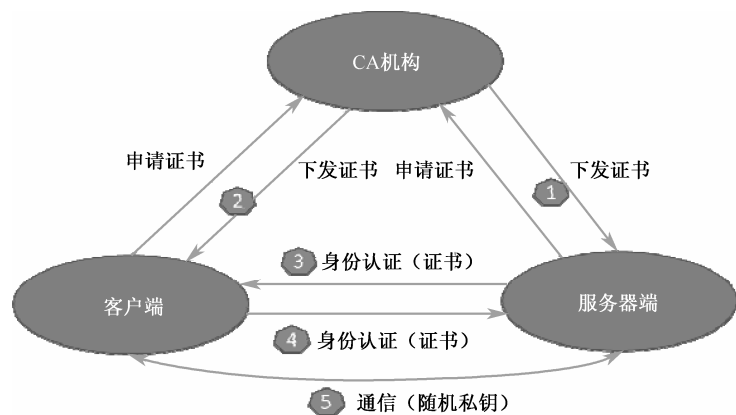


图 3.11 CA 认证流程

(1) HTTPS 通信双方的服务器端向 CA 机构申请证书，CA 机构是可信的第三方机构，它可以是一个公认的权威的企业，也可以是企业自身。企业内部系统一般都用企业自身的认证系统。CA 机构下发根证书、服务端证书及私钥给申请者。

(2) HTTPS 通信双方的客户端向 CA 机构申请证书，CA 机构下发根证书、客户端证书及私钥给申请者。

(3) 客户端向服务器端发起请求，服务端下发服务端证书给客户端。客户端接收到证书后，通过私钥解密证书，并利用服务器端证书中的公钥认证证书信息比较证书里的消息，例如域名和公钥与服务器刚刚发送的相关消息是否一致，如果一致，则客户端认可这个服务器的合法身份。

(4) 客户端发送客户端证书给服务器端，服务端接收到证书后，通过私钥解密证书，获得客户端证书公钥，并用该公钥认证证书信息，确认客户端是否合法。

(5) 客户端通过随机密钥加密信息，并发送加密后的信息给服务端。服务器端和客户端协商好加密方案后，客户端会产生一个随机的密钥，客户端通过协商好的加密方案，加密该随机

密钥，并发送该随机密钥到服务器端。服务器端接收这个密钥后，双方通信的所有内容都通过该随机密钥加密。

如上所述是双向认证 SSL 协议的具体通信过程，这种情况要求服务器和用户双方都有证书。单向认证 SSL 协议不需要客户拥有 CA 证书，对于上面的步骤，只需将服务器端验证客户证书的过程去掉，以及在协商对称密码方案 and 对称会话密钥时，服务器发送给客户的是没有加过密的（这并不影响 SSL 过程的安全性）密码方案。

其次，我们来看看 HTTP Token 的认证原理。

HTTP Token 的认证是用一个很长的特殊编码方式的并且难以被模仿的字符串——Token 来表明客户身份的一种方式。在通常情况下，Token 是一个很复杂的字符串，比如我们用私钥签名一个字符串后的数据就可以当作一个 Token。此外，每个 Token 对应一个用户名，存储在 API Server 能访问的一个文件中。当客户端发起 API 调用请求时，需要在 HTTP Header 里放入 Token，这样一来，API Server 就能识别合法用户和非法用户了。

最后，我们说说 HTTP Base 认证。

我们知道，HTTP 是无状态的，浏览器和 Web 服务器之间可以通过 Cookie 来进行身份识别。桌面应用程序（比如新浪桌面客户端、SkyDrive 客户端、命令行程序）一般不会使用 Cookie，那么它们与 Web 服务器之间是如何进行身份识别的呢？这就用到了 HTTP Base 认证，这种认证方式是把“用户名+冒号+密码”用 BASE64 算法进行编码后的字符串放在 HTTP Request 中的 Header Authorization 域里发送给服务端，服务端收到后进行解码，获取用户名及密码，然后进行用户身份的鉴权过程。

### 3.6.2 API Server 授权管理（Authorization）

---

当客户端发起 API Server 调用时，API Server 内部要先进行用户认证，然后执行用户授权流程，即通过“授权策略”来决定一个 API 调用是否合法。对合法用户进行授权（Authorization）并且随后在用户访问时进行鉴权，是权限与安全系统的重要一环。简单地说，授权就是授予不同的用户不同的访问权限，API Server 目前支持以下几种授权策略（通过 API Server 的启动参数“--authorization-mode”设置）。

- ◎ AlwaysDeny: 表示拒绝所有的请求，一般用于测试。
- ◎ AlwaysAllow: 允许接收所有请求，如果集群不需要授权流程，则可以采用该策略，这也是 Kubernetes 的默认配置。
- ◎ ABAC (Attribute-Based Access Control): 基于属性的访问控制，表示使用用户配置的授权规则对用户请求进行匹配和控制。

- ◎ **Webhook**：通过调用外部 REST 服务对用户进行授权。
- ◎ **RBAC**：Role-Based Access Control，基于角色的访问控制。

API Server 在接收到请求后，会读取该请求中的数据，生成一个访问策略对象，如果该请求中不带某些属性（如 Namespace），则这些属性的值将根据属性类型的不同，设置不同的默认值（例如为字符串类型的属性设置一个空字符串；为布尔类型的属性设置 false；为数值类型的属性设置 0）。然后将这个访问策略对象和授权策略文件中的所有访问策略对象逐条匹配，如果至少有一个策略对象被匹配，则该请求将被鉴权通过，否则终止 API 调用流程，并返回客户端的错误调用码。

## 1. ABAC 授权模式详解

在 API Server 启用 ABAC 模式时，需要指定授权策略文件的路径和名字（--authorization-policy-file=SOME\_FILENAME），授权策略文件里的每一行以一个 Map 类型的 JSON 对象进行设置，这被称为“访问策略对象”，通过设置“访问策略对象”中的下列属性来确定具体的授权策略。

- (1) apiVersion：当前版本为 abac.authorization.kubernetes.io/v1beta1。
- (2) kind：设置为“Policy”。
- (3) spec：详细的策略设置，包括下列字段。
  - (a) 主体属性
    - ◎ user（用户名）：字符串类型，该字符串类型的用户名来源于 Token 文件（--token-auth-file 参数设置的文件）或基本认证文件中用户名字段的值。
    - ◎ group（用户组）：设置为“system:authenticated”时表示匹配所有已认证的请求，设置为“system:unauthenticated”时表示匹配所有未认证的请求。
  - (b) 资源属性
    - ◎ apiGroup（API 组）：字符串类型，表明匹配哪些 API Group，例如 extensions 或\*（表示匹配所有 API Group）。
    - ◎ namespace（命名空间）：字符串类型，表明该策略允许访问某个 Namespace 的资源，例如 kube-system 或\*（表示匹配所有 namespace）。
    - ◎ resource（资源）：字符串类型，API 资源对象，例如 pods 或\*（表示匹配所有资源对象）。
  - (c) 非资源属性
    - ◎ nonResourcePath（非资源对象类路径）：非资源对象类的 URL 路径，例如/version 或/apis，



\*表示匹配所有非资源对象类的请求路径，也可以设置为子路径，/foo/\*表示匹配所有/foo 路径下的所有子路径。

- ◎ readonly（只读标识）：布尔类型，当它的值为 true 时，表明仅允许 GET 请求通过。

### 1) ABAC 授权算法

API Server 进行 ABAC 授权的算法如下：在 API Server 收到请求之后，这些请求携带的策略对象的属性就被组装为一个数组，接下来使用策略文件对其进行逐条匹配。如果有至少一行匹配成功，那么这个请求就通过了授权（不过还是可能会在后续其他授权校验中失败）。常见的策略配置如下。

- ◎ 要允许所有认证用户做某件事，可以写一个策略，将 group 属性设置为 system:authenticated。
- ◎ 要允许所有未认证用户做某件事，可以把策略的 group 属性设置为 system:unauthenticated。
- ◎ 要允许一个用户做任何事，将策略的 apiGroup、namespace、resource 和 nonResourcePath 属性设置为“\*”即可。

### 2) 使用 kubectl 时的授权机制

kubectl 使用 API Server 的/api 和/apis 端点来获取版本信息。要验证 kubectl create/update 命令发送给服务器的对象，kubectl 需要向 swagger 资源进行查询，Kubernetes v1 版的 API 对应的是/swaggerapi/api/v1 和/swaggerapi/experimental/v1。

当使用 ABAC 授权模式时，下列特殊资源必须显式地用 nonResourcePath 属性来表达。

- ◎ API 版本协商过程中的/api、/api/\*、/apis、和/apis/\*。
- ◎ 使用 kubectl version 命令从服务器获取版本时的/version。
- ◎ create/update 操作过程中的/swaggerapi/\*。

在使用 kubectl 操作时，如果需要查看发送到 API Server 的 HTTP 请求，则可以将日志级别设置为 8，例如：

```
# kubectl --v=8 version
```

下面通过几个授权策略文件（JSON 格式）示例说明 ABAC 的访问控制用法。

(1) 允许用户 alice 对所有资源做任何操作：

```
{ "apiVersion": "abac.authorization.kubernetes.io/v1beta1", "kind": "Policy",
  "spec": { "user": "alice", "namespace": "*", "resource": "*", "apiGroup": "*" } }
```

(2) kubelet 可以读取任意 Pod:

```
{"apiVersion": "abac.authorization.kubernetes.io/v1beta1", "kind": "Policy",  
"spec": {"user": "", "namespace": "*", "resource": "pods", "readOnly": true}}
```

(3) kubelet 可以读写 Event 对象:

```
{"apiVersion": "abac.authorization.kubernetes.io/v1beta1", "kind": "Policy",  
"spec": {"user": "kubelet", "namespace": "*", "resource": "events"}}
```

(4) 用户 bob 只能读取 projectCaribou 中的 Pod:

```
{"apiVersion": "abac.authorization.kubernetes.io/v1beta1", "kind": "Policy",  
"spec": {"user": "bob", "namespace": "projectCaribou", "resource": "pods",  
"readOnly": true}}
```

(5) 任何用户都可以对非资源类路径进行只读请求:

```
{"apiVersion": "abac.authorization.kubernetes.io/v1beta1", "kind": "Policy",  
"spec": {"group": "system:authenticated", "readOnly": true, "nonResourcePath": "*"}},  
{"apiVersion": "abac.authorization.kubernetes.io/v1beta1", "kind": "Policy",  
"spec": {"group": "system:unauthenticated", "readOnly": true, "nonResourcePath":  
"*"}}
```

如果添加了新的 ABAC 策略，则需要重启 API Server 以使其生效。

### 3) Service Account 与授权

Service Account 会自动生成一个用户 (user)，用户的名称按照以下规则产生：

```
system:serviceaccount:<namespace>:<serviceaccountname>
```

创建新的命名空间时，会产生一个如下形式的 Service Account：

```
system:serviceaccount:<namespace>:default
```

如果希望 kube-system 命名空间中的 Service Account “default” 具有全部权限，就要在策略文件中加入如下内容：

```
{"apiVersion": "abac.authorization.kubernetes.io/v1beta1", "kind": "Policy", "spec": {"user": "system:serviceaccount:kube-system:default", "namespace": "*", "resource": "*", "apiGroup": "*"}}
```

## 2. Webhook 授权模式详解

Webhook 定义了一个 HTTP 回调接口，实现 Webhook 的应用会在指定事件发生时，向一个 URL 地址 POST 通知信息。启用 Webhook 授权模式后，Kubernetes 会调用外部 REST 服务对用户进行授权。

Webhook 模式用参数 `--authorization-webhook-config-file=SOME_FILENAME` 来设置远端授权服务的信息。

配置文件使用的是 kubeconfig 文件的格式。文件里 **user** 一节的内容指的是 API Server。相对于远程授权服务来说，API Server 是客户端，也就是用户；**cluster** 一节的内容指的是远程授权服务器的配置。下面的例子为设置一个使用 HTTPS 客户端认证的配置：

```
clusters:           # clusters 指向远端服务
- name: name-of-remote-authz-service
  cluster:
    certificate-authority: /path/to/ca.pem      # 用于验证远端服务的 CA
    server: https://authz.example.com/authorize # 远端服务的 URL，必须使用 HTTPS

users:              # users 就是 API Server 的 Webhook 配置
- name: name-of-api-server
  user:
    client-certificate: /path/to/cert.pem # Webhook 插件使用的证书
    client-key: /path/to/key.pem         # 证书的 key
current-context: webhook      # kubeconfig 文件需要设置 context
contexts:
- context:
    cluster: name-of-remote-authz-service
    user: name-of-api-server
    name: webhook
```

在授权开始时，API Server 会生成一个 `api.authorization.v1beta1.SubjectAccessReview` 对象，用于描述操作信息，在进行 JSON 序列化之后 POST 出来。这个对象中包含了用户尝试访问资源的请求动作的描述，以及被访问资源的属性。

Webhook API 对象和其他 API 对象一样，遵循同样的版本兼容性规则，在实现时要注意 `apiVersion` 字段的版本，以实现正确的反序列化操作。另外，API Server 必须启用 `authorization.k8s.io/v1beta1` API 扩展（`--runtime-config=authorization.k8s.io/v1beta1=true`）。

下面是一个希望获取 Pod 列表的请求报文示例：

```
{
  "apiVersion": "authorization.k8s.io/v1beta1",
  "kind": "SubjectAccessReview",
  "spec": {
    "resourceAttributes": {
      "namespace": "kittensandponies",
      "verb": "get",
      "group": "unicorn.example.org",
      "resource": "pods"
    },
    "user": "jane",
    "group": [
      "group1",
      "group2"
    ]
  }
}
```

```
    ]  
  }  
}
```

远端服务需要填充请求中的 `SubjectAccessReviewStatus` 字段，并返回允许或不允许访问的结果。应答报文中的 `spec` 字段是无效的，也可以省略。

一个返回“运行访问”的应答报文示例如下：

```
{  
  "apiVersion": "authorization.k8s.io/v1beta1",  
  "kind": "SubjectAccessReview",  
  "status": {  
    "allowed": true  
  }  
}
```

一个返回“不允许访问”的应答报文示例如下：

```
{  
  "apiVersion": "authorization.k8s.io/v1beta1",  
  "kind": "SubjectAccessReview",  
  "status": {  
    "allowed": false,  
    "reason": "user does not have read access to the namespace"  
  }  
}
```

非资源的访问请求路径包括 `/api`、`/apis`、`/metrics`、`/resetMetrics`、`/logs`、`/debug`、`/healthz`、`/swagger-ui`、`/swaggerapi`、`/ui` 和 `/version`。通常可以对 `/api`、`/api/*`、`/apis`、`/apis/*` 和 `/version` 对于客户端发现服务器提供的资源和版本信息给予“允许”的授权，对于其他非资源的访问一般可以禁止，以限制客户端对 API Server 进行没有必要的查询。

查询 `/debug` 的请求报文示例如下：

```
{  
  "apiVersion": "authorization.k8s.io/v1beta1",  
  "kind": "SubjectAccessReview",  
  "spec": {  
    "nonResourceAttributes": {  
      "path": "/debug",  
      "verb": "get"  
    },  
    "user": "jane",  
    "group": [  
      "group1",  
    ]  
  }  
}
```

```

    "group2"
  ]
}
}

```

### 3. RBAC 授权模式详解

RBAC (Role-Based Access Control, 基于角色的访问控制) 在 Kubernetes v1.5 中引入, 在 v1.6 版本时升级为 Beta 版本, 并成为 kubeadm 安装方式下的默认选项, 足见其重要程度。相对于其他的访问控制方式, 新的 RBAC 具有如下优势。

- ◎ 对集群中的资源和非资源权限均有完整的覆盖。
- ◎ 整个 RBAC 完全由几个 API 对象完成, 同其他 API 对象一样, 可以用 kubectl 或 API 进行操作。
- ◎ 可以在运行时进行调整, 无须重新启动 API Server。

要使用 RBAC 授权模式, 则需要在 API Server 的启动参数中加上 `--authorization-mode=RBAC`。

下面对 RBAC 的原理和用法进行说明。

#### 1) RBAC 的 API 资源对象说明

RBAC 引入了 4 个新的顶级资源对象: Role、ClusterRole、RoleBinding 和 ClusterRoleBinding。同其他 API 资源对象一样, 用户可以使用 kubectl 或者 API 调用等方式操作这些资源对象。

##### ❖ 角色 (Role)

一个角色就是一组权限的集合, 这里的权限都是许可形式的, 不存在拒绝的规则。在一个命名空间中, 可以用角色来定义一个角色, 如果是集群级别的, 就需要使用 ClusterRole 了。

角色只能对命名空间内的资源进行授权, 下面例子中定义的角色具备读取 Pod 的权限:

```

kind: Role
apiVersion: rbac.authorization.k8s.io/v1beta1
metadata:
  namespace: default
  name: pod-reader
rules:
- apiGroups: [""]      # "" 空字符串, 表示核心 API 群
  resources: ["pods"]
  verbs: ["get", "watch", "list"]

```

rules 中的参数说明如下。

- ◎ apiGroups: 支持的 API 组列表, 例如 “apiVersion: batch/v1” “apiVersion: extensions:

v1beta1” “apiVersion: apps/v1beta1” 等，详细的 API 组说明参见第 4 章。

- ◎ **resources**：支持的资源对象列表，例如 pods、deployments、jobs 等。
- ◎ **verbs**：对资源对象的操作方法列表，例如 get、watch、list、delete、replace、patch 等，详细的操作方法说明参见第 4 章。

### ❖ 集群角色（ClusterRole）

集群角色除了具有和角色一致的命名空间内资源的管理能力，因其集群级别的范围，还可以用于以下特殊元素的授权。

- ◎ 集群范围的资源，例如 Node（节点）。
- ◎ 非资源型的路径，例如"/healthz"。
- ◎ 包含全部命名空间的资源，例如 pods（用于 kubectl get pods --all-namespaces 这样的操作授权）。

下面的集群角色可以让用户有权访问任意一个或所有命名空间的 secrets（视其绑定方式而定）：

```
kind: ClusterRole
apiVersion: rbac.authorization.k8s.io/v1beta1
metadata:
  # ClusterRole 不受限于命名空间，所以省略了 namespace name 的定义
rules:
- apiGroups: [""]
  resources: ["secrets"]
  verbs: ["get", "watch", "list"]
```

### ❖ 角色绑定（RoleBinding）和集群角色绑定（ClusterRoleBinding）

角色绑定或集群角色绑定用来把一个角色绑定到一个目标上，绑定目标可以是 User（用户）、Group（组）或者 Service Account。使用 RoleBinding 为某个命名空间授权，使用 ClusterRoleBinding 为集群范围内授权。

RoleBinding 可以引用 Role 进行授权。下例中的 RoleBinding 将在 default 命名空间中把 pod-reader 角色授予用户 jane，这一操作让 jane 可以读取 default 命名空间中的 Pod：

```
kind: RoleBinding
apiVersion: rbac.authorization.k8s.io/v1beta1
metadata:
  name: read-pods
  namespace: default
subjects:
- kind: User
  name: jane
```

```

  apiGroup: rbac.authorization.k8s.io
roleRef:
  kind: Role
  name: pod-reader
  apiGroup: rbac.authorization.k8s.io

```

**RoleBinding** 也可以引用 **ClusterRole**，对属于同一命名空间内 **ClusterRole** 定义的资源主体进行授权。一种常见的做法是集群管理员为集群范围预先定义好一组角色（**ClusterRole**），然后在多个命名空间中重复使用这些 **ClusterRole**。

例如下面的例子，虽然 **secret-reader** 是一个集群角色，但是因为使用了 **RoleBinding**，所以 **dave** 只能读取 **development** 命名空间中的 **secret**：

```

kind: RoleBinding
apiVersion: rbac.authorization.k8s.io/v1beta1
metadata:
  name: read-secrets
  namespace: development #集群角色中，只有在 development 命名空间中的权限才能赋予 dave
subjects:
- kind: User
  name: dave
  apiGroup: rbac.authorization.k8s.io
roleRef:
  kind: ClusterRole
  name: secret-reader
  apiGroup: rbac.authorization.k8s.io

```

集群角色绑定中的角色只能是集群角色，用于进行集群级别或者对所有命名空间都生效的授权。下面的例子允许 **manager** 组的用户读取任意 **namespace** 中的 **secret**：

```

kind: ClusterRoleBinding
apiVersion: rbac.authorization.k8s.io/v1beta1
metadata:
  name: read-secrets-global
subjects:
- kind: Group
  name: manager
  apiGroup: rbac.authorization.k8s.io
roleRef:
  kind: ClusterRole
  name: secret-reader
  apiGroup: rbac.authorization.k8s.io

```

图 3.12 展示了上述对 Pod 的 **get/watch/list** 操作进行授权的 **Role** 和 **RoleBinding** 逻辑关系。

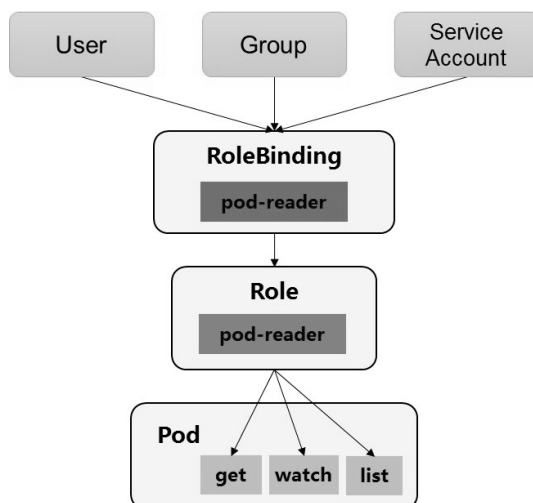


图 3.12 RoleBinding、Role 与对 Pod 的操作授权

## 2) 对资源的引用方式

多数资源可以用其名称的字符串来表达，也就是 Endpoint 中的 URL 相对路径，例如 pods。然而，某些 Kubernetes API 包含下级资源，例如 Pod 的日志(logs)。Pod 日志的 Endpoint 是 GET/api/v1/namespaces/{namespace}/pods/{name}/log。

在这个例子中，Pod 是一个命名空间内的资源，log 就是一个下级资源。要在一个 RBAC 角色中体现，则需要用斜线“/”来分隔资源和下级资源。若想授权让某个主体同时能够读取 Pod 和 Pod log，则可以配置 resources 为一个数组：

```
kind: Role
apiVersion: rbac.authorization.k8s.io/v1beta1
metadata:
  namespace: default
  name: pod-and-pod-logs-reader
rules:
- apiGroups: [""]
  resources: ["pods", "pods/log"]
  verbs: ["get", "list"]
```

资源还可以通过名字(ResourceName)进行引用。在指定 ResourceName 后，使用 get、delete、update 和 patch 动词的请求，就会被限制在这个资源实例范围内。例如下面的声明让一个主体只能对一个 configmap 进行 get 和 update 操作：

```
kind: Role
apiVersion: rbac.authorization.k8s.io/v1beta1
metadata:
```



```

namespace: default
name: configmap-updater
rules:
- apiGroups: [""]
  resources: ["configmap"]
  resourceNames: ["my-configmap"]
  verbs: ["update", "get"]

```

可想而知，resourceName 这种用法对 list、watch、create 或 deletecollection 操作是无效的，这是因为必须要通过 URL 进行鉴权，而资源名称在 list、watch、create 或 deletecollection 请求中只是请求 Body 数据的一部分。

下面对常见的角色（Role）和角色绑定（RoleBinding）给出示例，提供参考用法。

### 3）常用的角色（Role）示例

注意，下面的例子只展示了 rules 部分的内容。

（1）允许读取核心 API 组中的 Pod 资源：

```

rules:
- apiGroups: [""]
  resources: ["pods"]
  verbs: ["get", "list", "watch"]

```

（2）允许读写"extensions"和"apps"两个 API 组中的"deployment"资源：

```

rules:
- apiGroups: ["extensions", "apps"]
  resources: ["deployments"]
  verbs: ["get", "list", "watch", "create", "update", "patch", "delete"]

```

（3）允许读取"pods"及读写"jobs"：

```

rules:
- apiGroups: [""]
  resources: ["pods"]
  verbs: ["get", "list", "watch"]
- apiGroups: ["batch", "extensions"]
  resources: ["jobs"]
  verbs: ["get", "list", "watch", "create", "update", "patch", "delete"]

```

（4）允许读取一个名为"my-config"的 ConfigMap（必须绑定到一个 RoleBinding 来限制到一个 namespace 下的 ConfigMap）：

```

rules:
- apiGroups: [""]
  resources: ["configmaps"]
  resourceNames: ["my-config"]
  verbs: ["get"]

```

（5）读取核心组的"node"资源（Node 属于集群级的资源，所以必须存在于 ClusterRole 中，并使用 ClusterRoleBinding 进行绑定）：

```
rules:
- apiGroups: [""]
  resources: ["nodes"]
  verbs: ["get", "list", "watch"]
```

（6）允许对非资源端点"/healthz"及其所有子路径进行"GET"和"POST"操作（必须使用 ClusterRole 和 ClusterRoleBinding）：

```
rules:
- nonResourceURLs: ["/healthz", "/healthz/*"]
  verbs: ["get", "post"]
```

#### 4）常用的角色绑定（RoleBinding）示例

注意，下面的例子中只包含 subjects 部分的内容。

（1）用户名"alice@example.com"：

```
subjects:
- kind: User
  name: "alice@example.com"
  apiGroup: rbac.authorization.k8s.io
```

（2）组名"frontend-admins"：

```
subjects:
- kind: Group
  name: "frontend-admins"
  apiGroup: rbac.authorization.k8s.io
```

（3）kube-system 命名空间中的默认 Service Account：

```
subjects:
- kind: ServiceAccount
  name: default
  namespace: kube-system
```

（4）"qa"命名空间中的所有 Service Account：

```
subjects:
- kind: Group
  name: system:serviceaccounts:qa
  apiGroup: rbac.authorization.k8s.io
```

（5）所有 Service Account：

```
subjects:
- kind: Group
  name: system:serviceaccounts
  apiGroup: rbac.authorization.k8s.io
```

(6) 所有认证用户 (v1.5 版本以上):

```
subjects:
- kind: Group
  name: system:authenticated
  apiGroup: rbac.authorization.k8s.io
```

(7) 所有未认证用户 (v1.5 版本以上):

```
subjects:
- kind: Group
  name: system:unauthenticated
  apiGroup: rbac.authorization.k8s.io
```

(8) 全部用户 (v1.5 版本以上):

```
subjects:
- kind: Group
  name: system:authenticated
  apiGroup: rbac.authorization.k8s.io
- kind: Group
  name: system:unauthenticated
  apiGroup: rbac.authorization.k8s.io
```

## 5) 默认的角色和角色绑定

API Server 会创建一套默认的 ClusterRole 和 ClusterRoleBinding 对象，其中很多是以“system:”为前缀的，以表明这些资源属于基础架构，对这些对象的改动可能造成集群故障。举例来说，system:node 这个 ClusterRole 为 kubelet 定义了权限，如果这个集群角色被改动了，则 kubelet 会停止工作。

所有默认的 ClusterRole 和 RoleBinding 都会用标签 `kubernetes.io/bootstrapping=rbac-defaults` 进行标记。

下面对一些常见的默认 ClusterRole 和 ClusterRoleBinding 对象进行说明。

对系统角色的说明如表 3.1 所示。

表 3.1 系统角色

默认的 ClusterRole	默认的 ClusterRoleBinding	描述
system:basic-user	system:authenticated 和 system:unauthenticated 组	让用户能够读取自身的信息
system:discovery	system:authenticated 和 system:unauthenticated 组	对 API 发现 Endpoint 的只读访问，用于 API 级别的发现和协商

对用户角色的说明如表 3.2 所示。

有些默认角色不是以“system:”为前缀的，这部分角色是针对用户的。其中包含超级用户

角色（cluster-admin），有的用于集群一级的角色（cluster-status），还有针对 namespace 的角色（admin,edit,view）。

表 3.2 用户角色

默认的 ClusterRole	默认的 ClusterRoleBinding	描 述
cluster-admin	system:masters 组	让超级用户可以对任何资源执行任何操作。如果在 ClusterRoleBinding 中使用，则影响的是整个集群的所有 namespace 中的任何资源；如果使用的是 RoleBinding，则能控制这一绑定的 namespace 中的资源，还包括 namespace 本身
cluster-status	None	可以对基础集群状态信息进行只读访问
admin	None	允许 admin 访问，可以限制在一个 namespace 中使用 RoleBinding。如果在 RoleBinding 中使用，则允许对 namespace 中的大多数资源进行读写访问，其中包含创建角色和角色绑定的能力。这一角色不允许操作 namespace 本身，也不能写入资源限额
edit	None	允许对命名空间内的大多数资源进行读写操作，不允许查看或修改角色，以及角色绑定
view	None	允许对多数对象进行只读操作，但是对角色、角色绑定及 secret 是不可访问的

对核心 Master 组件角色的说明如表 3.3 所示。

表 3.3 核心 Master 组件角色

默认的 ClusterRole	默认的 ClusterRoleBinding	描 述
system:kube-scheduler	system:kube-scheduler 用户	能够访问 kube-scheduler 组件所需的资源
system:kube-controller-manager	system:kube-controller-manager 用户	能够访问 kube-controller-manager 组件所需的资源。不同的控制所需的不同权限参见表 3.4
system:node	system:nodes 组	允许访问 kubelet 所需的资源，包括对 secret 的读取，以及对 Pod 的写入。未来会把上面的两个权限限制在分配到本 Node 的对象上。今后的鉴权过程，kubelet 必须以 system:node 及一个 system:node 形式的用户名进行。参看 <a href="https://pr.k8s.io/40476">https://pr.k8s.io/40476</a>
system:node-proxier	system:kube-proxy 用户	允许访问 kube-proxy 所需的资源
system:kube-scheduler	system:kube-scheduler 用户	能够访问 kube-scheduler 组件所需的资源

对其他组件角色的说明如表 3.4 所示。

表 3.4 其他组件角色

默认的 ClusterRole	默认的 ClusterRoleBinding	描 述
system:auth-delegator	None	允许对授权和认证进行托管，通常用于附加的 API 服务器
system:heapster	None	Heapster 组件的角色
system:kube-aggregator	None	kube-aggregator 的角色

续表

默认的 ClusterRole	默认的 ClusterRoleBinding	描 述
system:kube-dns	在 kube-system namespace 中 kube-dns 的 Service Account	kube-dns 的角色
system:node-bootstrapper	None	允许访问 kubelet TLS 启动所需的资源
system:node-problem-detector	None	允许访问 node-problem-detector 组件所需的资源
system:persistent-volume-provisioner	None	允许访问多数动态卷供给所需的资源

对 Controller 角色的说明如表 3.5 所示。

表 3.5 Controller 角色

需要赋予的角色
system:controller:attachdetach-controller
system:controller:certificate-controller
system:controller:cronjob-controller
system:controller:daemon-set-controller
system:controller:deployment-controller
system:controller:disruption-controller
system:controller:endpoint-controller
system:controller:generic-garbage-collector
system:controller:horizontal-pod-autoscaler
system:controller:job-controller
system:controller:namespace-controller
system:controller:node-controller
system:controller:persistent-volume-binder
system:controller:pod-garbage-collector
system:controller:replicaset-controller
system:controller:replication-controller
system:controller:resourcequota-controller
system:controller:route-controller
system:controller:service-account-controller
system:controller:service-controller
system:controller:statefulset-controller
system:controller:ttl-controller

Kubernetes Controller Manager 负责的是核心控制流。如果用--use-service-account-credentials调用，则每个控制过程都会使用不同的 Service Account 启动，因此就有了对应各个控制过程的

角色，前缀是 `system:controller`。如果 `Controller Manager` 没有用 `--use-service-account-credentials` 启动参数，则将使用自己的凭据运行各个控制流程，这就需要为该凭据授予所有相关角色。

## 6) 授权注意事项：预防提权和授权初始化

RBAC API 拒绝用户利用编辑角色或者角色绑定的方式进行提权。这一限制是在 API 层面做出的，因此即使 RBAC 没有启用也仍然有效。

用户只能在拥有一个角色的所有权限，且与该角色的生效范围一致（如果是集群角色，则是集群范围；如果是普通角色，则可能是同一个命名空间或者全集群）的前提下，才能对角色进行创建和更新。例如用户 `user-1` 没有列出集群中所有 `secret` 的权限，就不能创建具有这一权限的集群角色。要让一个用户能够创建或更新角色，需要：

- ◎ 为其授予一个允许创建/更新 `Role` 或 `ClusterRole` 资源对象的角色；
- ◎ 为用户授予角色，要覆盖该用户所能控制的所有权限范围。用户如果尝试创建超出其自身权限的角色或者集群角色，则该 API 调用会被禁止。

如果一个用户的权限包含了一个角色的所有权限，那么就可以为其创建和更新角色绑定（要求同样的作用范围）；或者如果被授予了针对某个角色的绑定授权，则也有权完成此操作。例如如果用户 `user-1` 没有列出集群内所有 `secret` 的权限，就无法为一个具有这样权限的角色创建集群角色绑定。要使用户能够创建、更新这一角色绑定，则需要有如下做法。

(1) 为其授予一个允许创建和更新角色绑定或者集群角色绑定的角色。

(2) 为其授予绑定某一角色的权限，有隐式或显式两种方法。

- ◎ 隐式：让其具有所有该角色的权限。
- ◎ 显式：为用户授予针对该角色（或集群角色）的绑定操作的权限。

例如，下面的集群角色和角色绑定能让 `user-1` 为其他用户在 `user-1-namespace` 命名空间中授予 `admin`、`edit` 及 `view` 角色：

```
apiVersion: rbac.authorization.k8s.io/v1beta1
kind: ClusterRole
metadata:
  name: role-grantor
rules:
- apiGroups: ["rbac.authorization.k8s.io"]
  resources: ["rolebindings"]
  verbs: ["create"]
- apiGroups: ["rbac.authorization.k8s.io"]
  resources: ["clusterroles"]
  verbs: ["bind"]
  resourceNames: ["admin", "edit", "view"]
```

```

---
apiVersion: rbac.authorization.k8s.io/v1beta1
kind: RoleBinding
metadata:
  name: role-grantor-binding
  namespace: user-1-namespace
roleRef:
  apiGroup: rbac.authorization.k8s.io
  kind: ClusterRole
  name: role-grantor
subjects:
- apiGroup: rbac.authorization.k8s.io
  kind: User
  name: user-1

```

在进行第 1 个角色和角色绑定时，必须让初始用户具备其尚未被授予的权限，要进行初始的角色和角色绑定设置，有以下两种办法。

(1) 使用属于 `system:masters` 组的身份，这一群组默认具有 `cluster-admin` 这一超级角色的绑定。

(2) 如果 API Server 以 `--insecure-port` 参数运行，则客户端通过这个非安全端口进行接口调用，这一端口没有认证鉴权的限制。

### 7) 对 Service Account 的授权管理

默认的 RBAC 策略为控制平台组件、节点和控制器授予有限范围的权限，但是在 "kube-system" 之外的 Service Account 是没有任何权限的（除了所有认证用户都具有的 `discovery` 权限）。

这就要求用户为 Service Account 赋予所需的权限。细粒度的角色分配能够提高安全性，但也会提高管理成本。粗放的授权方式可能会给 Service Account 多余的权限，但会更易于管理。

下面的实践以安全性递减的方式排序。

(1) 为一个应用专属的 Service Account 赋权（最佳实践）。

这个应用需要在 Pod 的 Spec 中指定一个 `serviceAccountName`，用 `API`、`Application Manifest`、`kubectl create serviceaccount` 命令等创建 Service Account，例如为 "my-namespace" 中的 "my-sa" Service Account 授予只读权限：

```

$ kubectl create rolebinding my-sa-view \
  --clusterrole=view \
  --serviceaccount=my-namespace:my-sa \
  --namespace=my-namespace

```

（2）为一个命名空间中的"default" Service Account 授权。

如果一个应用没有指定 serviceAccountName，则会使用"default" Service Account。注意，赋给"default" Service Account 的权限会让所有没指定 serviceAccountName 的 Pod 都具有这些权限。

例如在"my-namespace"命名空间里为"default" Service Account 授予只读权限：

```
$ kubectl create rolebinding default-view \
  --clusterrole=view \
  --serviceaccount=my-namespace:default \
  --namespace=my-namespace
```

目前不少 Add-Ons 在"kube-system"命名空间中用"default" Service Account 运行。要让这些 Add-Ons 能够使用超级用户权限，则可以把 cluster-admin 权限赋予"kube-system"的"default" Service Account。注意，这一操作意味着"kube-system"命名空间包含了通向 API 超级用户的捷径：

```
$ kubectl create clusterrolebinding add-on-cluster-admin \
  --clusterrole=cluster-admin \
  --serviceaccount=kube-system:default
```

（3）为命名空间内的所有 Service Account 授予一个角色。

如果希望在一个命名空间里，任何 Service Account 的应用都具有一个角色，则可以为这一命名空间的 Service Account 群组进行授权。

例如，为"my-namespace"命名空间中的所有 Service Account 赋予只读权限：

```
$ kubectl create rolebinding serviceaccounts-view \
  --clusterrole=view \
  --group=system:serviceaccounts:my-namespace \
  --namespace=my-namespace
```

（4）为集群范围内的所有 Service Account 授予一个低权限角色（不推荐）。

如果不想为每个命名空间管理授权，则可以把一个集群级别的角色赋给所有 Service Account。

例如，为所有命名空间中的所有 Service Account 授予只读权限：

```
$ kubectl create clusterrolebinding serviceaccounts-view \
  --clusterrole=view \
  --group=system:serviceaccounts
```

（5）为所有 Service Account 授予超级用户权限（强烈不建议这样设置）。

如果完全不在意权限，则可以把超级用户权限分配给每个 Service Account。

注意，这会让所有具有读取 Secret 权限的用户都可以创建 Pod 来访问超级用户的专属权限：

```
$ kubectl create clusterrolebinding serviceaccounts-cluster-admin \
  --clusterrole=cluster-admin \
  --group=system:serviceaccounts
```



## 8) 使用 kubectl 命令行工具创建资源对象

除了使用 yaml 配置文件来创建这些资源对象，也可以直接使用 kubectl 命令行工具对它们进行创建。下面通过几个例子进行说明。

(1) 在命名空间 `acme` 内为用户 `bob` 授权 `admin ClusterRole`:

```
kubectl create rolebinding bob-admin-binding --clusterrole=admin --user=bob
--namespace=acme
```

(2) 在命名空间 `acme` 内为名为 `myapp` 的 Service Account 授予 `view ClusterRole`:

```
kubectl create rolebinding myapp-view-binding --clusterrole=view
--serviceaccount=acme:myapp --namespace=acme
```

(3) 在全集群范围内为用户 `"root"` 授予 `cluster-admin ClusterRole`:

```
kubectl create clusterrolebinding root-cluster-admin-binding
--clusterrole=cluster-admin --user=root
```

(4) 在全集群范围内为用户 `"kubelet"` 授予 `system:node ClusterRole`:

```
kubectl create clusterrolebinding kubelet-node-binding
--clusterrole=system:node --user=kubelet
```

(5) 在全集群范围内为名为 `"myapp"` 的 Service Account 授予 `view ClusterRole`:

```
kubectl create clusterrolebinding myapp-view-binding --clusterrole=view
--serviceaccount=acme:myapp
```

可以在 `kubectl --help` 的帮助中查看更详细的说明。

## 9) RBAC 的 Auto-reconciliation (自动恢复) 功能

自动恢复从 Kubernetes v1.6 版本开始引入。每次启动时，API Server 都会更新默认的集群角色的缺失权限，也会刷新默认的角色绑定中缺失的主体，这样就防止了一些破坏性的修改，也保证了在集群升级的情况下相关内容能够及时更新。

如果不希望使用这一功能，则可以将一个默认的集群角色（ClusterRole）或者角色绑定（RoleBinding）的 Annotation 注解 `"rbac.authorization.kubernetes.io/autoupdate"` 值设置为 `false`。

## 10) 从旧版本的授权策略升级到 RBAC

在 Kubernetes v1.6 之前，很多 Deployment 都试用了比较宽松的 ABAC 策略，包含为所有 Service Account 开放完全 API 访问。

默认的 RBAC 策略为控制台组件、节点和控制器授予了范围受限的权限，但是不会为 `"kube-system"` 以外的 Service Account 授予任何权限。

这样一来，可能会对现有的一些工作负载造成影响，这时有两种办法来解决这一问题。

(1) 并行认证。RBAC 和 ABAC 同时运行，并包含传统的 ABAC 策略：

```
--authorization-mode=RBAC,ABAC --authorization-policy-file=mypolicy.jsonl
```

首先会由 RBAC 尝试对请求进行鉴权，如果得到的结果是拒绝，那么就轮到 ABAC 生效。这样所有应用只要满足 RBAC 或 ABAC 之一即可工作。

如果使用 2 或者更高的日志标准（--v=2），则将可以在 API Server 日志中看到 RBAC 的拒绝行为（前缀：RBAC DENY）。可以利用这一信息来确定需要授予何种权限给用户、组或 Service Account。如果有一天，集群管理员已经按照 RBAC 的方式对 Service 进行了授权，并且这些工作负载运行的过程中不再出现 RBAC 的拒绝信息，就可以移除 RBAC 了。

（2）粗放管理。可以使用 RBAC 的角色绑定，复制一个粗放的策略。

警告：下面的策略让所有 Service Account 都具备了集群管理员权限，所有容器运行的应用都会自动接收到 Service Account 的认证，能够对任何 API 做任何事情，包括查看 Secret 和修改授权。这不是一个值得推荐的策略。

```
$ kubectl create clusterrolebinding permissive-binding \
  --clusterrole=cluster-admin \
  --user=admin \
  --user=kubelet \
  --group=system:serviceaccounts
```

### 3.6.3 Admission Control（准入控制）

突破了之前所说的认证和鉴权两道关口之后，客户端的调用请求就能够得到 API Server 的真正响应了吗？答案是：不能！这个请求还需要通过 Admission Control 所控制的一个“准入控制链”的层层考验，官方标准的“关卡”有近十个之多，而且能自定义扩展！笔者忽然在想，如果在幼儿园时，老师就告诉我们长大后还要读小学，参加中考、高考、公司面试、职称考试，等等，我们还会天天去幼儿园吗？

Admission Control 配备有一个“准入控制器”的插件列表，发送给 API Server 的任何请求都需要通过列表中每个准入控制器的检查，检查不通过，则 API Server 拒绝此调用请求。此外，准入控制器插件还能够修改请求参数以完成一些自动化的任务，比如 ServiceAccount 这个控制器插件。当前可配置的准入控制器插件如下。

- ◎ AlwaysAdmit：允许所有请求。
- ◎ AlwaysPullImages：在启动容器之前总是尝试重新下载镜像。这对于多租户共享一个集群的场景非常有用，系统在启动容器之前可以保证总是使用租户的密钥去下载镜像。如果不设置这个控制器，则在 Node 上下载的镜像的安全性将被削弱，只要知道该镜像的名称，任何人便可以使用它们了。

- ◎ **AlwaysDeny**: 禁止所有请求，用于测试。
- ◎ **DenyExecOnPrivileged**: 已弃用，拦截所有想在 **Privileged Container** 上执行命令的请求。如果你的集群支持 **Privileged Container**，又希望限制用户在这些 **Privileged Container** 上执行命令，那么强烈推荐你使用它。其功能已合并到 **DenyEscalatingExec** 中。
- ◎ **DenyEscalatingExec**: 拦截所有 **exec** 和 **attach** 到具有特权的 **Pod** 上的请求。如果你的集群支持运行有 **escalated privilege** 权限的容器，又希望限制用户在这些容器内执行命令，那么强烈推荐你使用它。
- ◎ **ImagePolicyWebhook**: 这个插件将允许后端的一个 **Webhook** 程序来完成 **admission controller** 的功能。**ImagePolicyWebhook** 需要使用一个配置文件（通过 **kube-apiserver** 的启动参数 `--admission-control-config-file` 设置）定义后端 **Webhook** 的参数。目前是 **Alpha** 版本的功能。
- ◎ **ServiceAccount**: 这个插件将 **ServiceAccount** 实现了自动化，如果你想要使用 **ServiceAccount** 对象，那么强烈推荐你使用它，在后面讲述 **ServiceAccount** 的章节会详细说明其作用。
- ◎ **SecurityContextDeny**: 这个插件将 **Pod** 中定义的 **SecurityContext** 选项全部失效。**SecurityContext** 在 **Container** 中定义了操作系统级别的安全设定（**uid**、**gid**、**capabilities**、**SELinux** 等）。在未设置 **PodSecurityPolicy** 的集群中建议启用该插件，以禁用容器设置的非安全访问权限。
- ◎ **ResourceQuota**: 用于资源配额管理目的，作用于 **Namespace** 上。该插件拦截所有请求，以确保在 **Namespace** 上的资源配额使用不会超标。推荐在 **Admission Control** 参数列表中将这个插件排最后一个，以免可能被其他插件拒绝的 **Pod** 被过早分配资源。在 5.1.4 节将详细介绍 **ResourceQuota** 的原理和用法。
- ◎ **LimitRanger**: 用于资源限制管理，作用于 **Namespace** 上，确保对 **Pod** 进行资源限制。启用该插件还会为未设置资源限制的 **Pod** 进行默认设置，例如为 **namespace “default”** 中的所有 **Pod** 设置 **0.1CPU** 的资源请求。在 5.1.4 节将详细介绍 **LimitRange** 的原理和用法。
- ◎ **InitialResources**: 为实验性特性，是一个开发中的功能，旨在为未设置资源请求、限制的 **Pod**，根据其镜像的历史资源的使用情况进行初始化的资源请求、限制设置。
- ◎ **NamespaceLifecycle**: 如果尝试在一个不存在的 **namespace** 中创建资源对象，则该创建请求将被拒绝。当删除一个 **namespace** 时，系统将会删除该 **namespace** 中的所有对象，包括 **Pod**、**Service** 等。

- ◎ **DefaultStorageClass**: 为了实现共享存储的动态供应，为未指定 **StorageClass** 或 **PV** 的 **PVC** 尝试匹配默认的 **StorageClass**，尽可能减少用户在申请 **PVC** 时所需了解的后端存储细节。关于 **StorageClass**、**PV**、**PVC** 的原理和用法详见 3.8 节的说明。
- ◎ **DefaultTolerationSeconds**: 这个插件为那些没有设置 **forgiveness tolerations** 并具有 **notready:NoExecute** 和 **unreachable:NoExecute** 两种 **taints** 的 **Pod** 设置默认的“容忍”时间，为 5min。
- ◎ **PodSecurityPolicy**: 这个插件用于在创建或修改 **Pod** 时决定是否根据 **Pod** 的 **security context** 和可用的 **PodSecurityPolicy** 对 **Pod** 的安全策略进行控制。

在 **API Server** 上设置 **--admission-control** 参数，即可定制我们需要的准入控制链，如果启用多种准入控制选项，则建议的设置（含加载顺序）如下。

对 **Kubernetes v1.6.0** 及以上版本设置如下：

```
--admission-control=NamespaceLifecycle,LimitRanger,ServiceAccount,PersistentVolumeLabel,DefaultStorageClass,ResourceQuota,DefaultTolerationSeconds
```

对 **Kubernetes >= v1.4.0** 版本：

```
--admission-control=NamespaceLifecycle,LimitRanger,ServiceAccount,DefaultStorageClass,ResourceQuota
```

对 **Kubernetes >= v1.2.0** 版本：

```
--admission-control=NamespaceLifecycle,LimitRanger,ServiceAccount,ResourceQuota
```

对 **Kubernetes >= v1.0.0** 版本：

```
--admission-control=NamespaceLifecycle,LimitRanger,SecurityContextDeny,ServiceAccount,PersistentVolumeLabel,ResourceQuota
```

### 3.6.4 Service Account

**Service Account** 也是一种账号，但它并不是给 **Kubernetes** 集群的用户（系统管理员、运维人员、租户用户等）使用的，而是给运行在 **Pod** 里的进程用的，它为 **Pod** 里的进程提供必要的身份证明。

在继续学习之前，请回忆一下本章前面所说的 **API Server** 的认证一节。

我们知道，正常情况下，为了确保 **Kubernetes** 集群的安全，**API Server** 都会对客户端进行身份认证，认证失败的客户端无法进行 **API** 调用。此外，**Pod** 中访问 **Kubernetes API Server** 服务时，是以 **Service** 方式访问服务名为 **Kubernetes** 的这个服务的，而 **Kubernetes** 服务又只在 **HTTPS** 安全端口 443 上提供服务，那么如何进行身份认证呢？这的确是个谜，因为 **Kubernetes** 的官方

文档并没有清楚说明这个问题。

通过查看官方源码，我们发现这是在用一种类似 HTTP Token 的新的认证方式——Service Account Auth，Pod 中的客户端调用 Kubernetes API 时，在 HTTP Header 中传递了一个 Token 字符串，这类似于之前提到的 HTTP Token 认证方式，但又有以下几个不同点。

- ◎ 这个 Token 的内容来自 Pod 里指定路径下的一个文件（/run/secrets/kubernetes.io/serviceaccount/token），这种 Token 是动态生成的，确切地说，是由 Kubernetes Controller 进程用 API Server 的私钥（--service-account-private-key-file 指定的私钥）签名生成的一个 JWT Secret。
- ◎ 官方提供的客户端 REST 框架代码里，通过 HTTPS 方式与 API Server 建立连接后，会用 Pod 里指定路径下的一个 CA 证书（/run/secrets/kubernetes.io/serviceaccount/ca.crt）验证 API Server 发来的证书，验证是否是被 CA 证书签名的合法证书。
- ◎ API Server 收到这个 Token 以后，采用自己的私钥（实际是使用参数 service-account-key-file 指定的私钥，如果此参数没有设置，则默认采用 tls-private-key-file 指定的参数，即自己的私钥）对 Token 进行合法性验证。

明白了认证原理，我们接下来继续分析上面认证过程中所涉及的 Pod 中的以下三个文件。

- ◎ /run/secrets/kubernetes.io/serviceaccount/token。
- ◎ /run/secrets/kubernetes.io/serviceaccount/ca.crt。
- ◎ /run/secrets/kubernetes.io/serviceaccount/namespace （客户端采用这里指定的 namespace 作为参数调用 Kubernetes API）。

这三个文件由于参与到 Pod 进程与 API Server 认证的过程中，起到了类似 Secret（私密凭据）的作用，所以它们被称为 Kubernetes Secret 对象。Secret 从属于 Service Account 资源对象，属于 Service Account 的一部分，一个 Service Account 对象里面可以包括多个不同的 Secret 对象，分别用于不同目的认证活动。

下面我们通过运行一些命令来加深我们对 Service Account 与 Secret 的直观认识。

首先，查看系统中的 Service Account 对象，我们看到有一个名为 default 的 Service Account 对象，包含一个名为 default-token-77oyg 的 Secret，这个 Secret 同时是“Mountable secrets”，表明它是需要被 Mount 到 Pod 上的：

```
# kubectl describe serviceaccounts
Name:         default
Namespace:    default
Labels:       <none>
Image pull secrets: <none>
```

```
# kubectl describe secrets default-token-77oyg
Name:         default-token-77oyg
Namespace:    default
Labels:       <none>
Annotations:  kubernetes.io/service-account.name=default
              kubernetes.io/service-account.uid=3e5b99c0-432c-11e6-b45c-000c29dc2102
```

```
Data
====
token:
eyJhbGciOiJSUzI1NiIsInR5cCI6IkpXVCJ9.eyJpc3MiOiJrdWJlcm5ldGVzL3NlcnZpY2VhY2NvIiwia3ViZXJzZXRLcy5pbpy9zZXJ2aWNlYWYjY2VudC9uZWllc3BhY2UiOiJkZWZhZDw0Iiwia3ViZXRLcy5pbpy9zZXJ2aWNlYWYjY2VudC9zZWYyZSIsImRlZmF1bHQtZGF9rZW4tNzdveWciLCJrdWJlcm5ldGVzLmlvL3NlcnZpY2VhY2NvdW50L3NlcnZpY2UtYWNjb3VudC5uZWllIjoieGVMiXV3Smt1YmVybmV0ZXMuaW8vc2VydmljZWZjY291bnQvc2VydmljZS1hY2NvdW50LnVpZCI6IjNlNWl5LTQzMmMtMTFlNiilINDVjLTAwMGMyOWRjMjEwMmEiIsInN1YiI6ImN5c3RlbTgzZXJ2aWNlYWYjY2VudC9zZWZhZDw0O0mRlZmF1bHQtZGF9rMFsBrYmTLMB55X3UGfO_pADP6FSsQgHb0SxGjTtsJnY-ze2vfFc8QdVdmQfFbnkgLWhltKIPr_EyvJTRP538uovgcA_QGN9yIMEdqIfQC2wfnLFuk10a8OdSH4uzayBb50jGWXWbXn6u0wAGMneiTKtCvzGFR4q-p19Jjh5qNPiUdJ0NhjsJSAclhdNK40XtOgMHdNNYPemPgk2cm7DRb6ifiSOs05cTeLYv1TpIBMvcQy4sYedCEL2cJ20BwcSo4-1Dev9rdxr5OdtgCvo60xbPF7RjjgUMLYO3YCi07WmqNdmxWHJkwvBtkWZhzhdvuFCpHeWANA
```

从上面的输出信息中我们看到，`default-token-77oyg` 包括三个数据项，分别是 `token`、`ca.crt`、`namespace`。联想到“Mountable secrets”的标记，以及之前看到的 Pod 中的三个文件的文件名，你可能恍然大悟，原来是这么一回事：每个 Namespace 下有一个名为 `default` 的默认的 Service Account 对象，这个 Service Account 里面有一个名为 `Tokens` 的可以当作 Volume 一样被 Mount 到 Pod 里的 Secret，当 Pod 启动时，这个 Secret 会自动被 Mount 到 Pod 的指定目录下，用来协助完成 Pod 中的进程访问 API Server 时的身份鉴权过程。

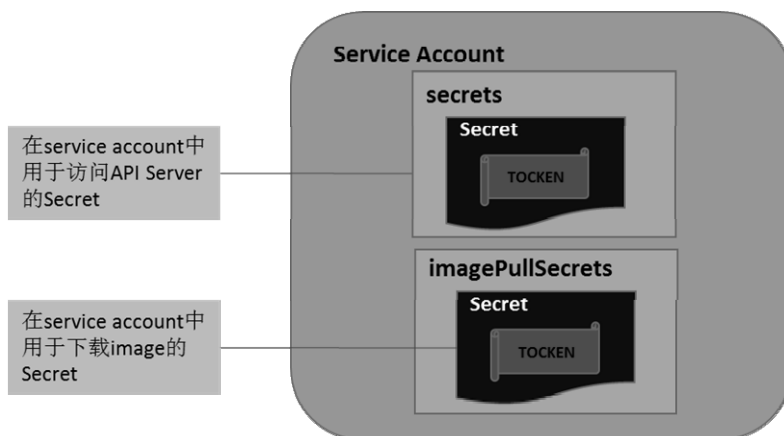


图 3.13 Service Account 中的 Secret

(1) 名为 Tokens 的 Secret 用于访问 API Server 的 Secret，也被称为 Service Account Secret。

(2) 名为 Image pull secrets 的 Secret 用于下载容器镜像时的认证过程，通常镜像库运行在 Insecure 模式下，所以这个 Secret 为空。

(3) 用户自定义的其他 Secret，用于用户的进程。

如果一个 Pod 在定义时没有指定 `spec.serviceAccountName` 属性，则系统会自动为其赋值为 “default”，即大家都使用同一个 Namespace 下默认的 Service Account。如果某个 Pod 需要使用非 default 的 Service Account，则需要在定义时指定：

```
apiVersion: v1
kind: Pod
metadata:
  name: mypod
spec:
  containers:
    - name: mycontainer
      image: nginx:v1
  serviceAccountName: myserviceaccount
```

Kubernetes 之所以要创建两套独立的账号系统，原因如下。

- ◎ User 账号是给人用的，Service Account 是给 Pod 里的进程使用的，面向的对象不同。
- ◎ User 账号是全局性的，Service Account 则属于某个具体的 Namespace。
- ◎ 通常来说，User 账号是与后端的用户数据库同步的，创建一个新用户通常要走一套复杂的业务流程才能实现，Service Account 的创建则需要极轻量级的实现方式，集群管理员可以很容易为某些特定任务创建一个 Service Account。

- ◎ 对于这两种不同的账号，其审计要求通常不同。
- ◎ 对于一个复杂的系统来说，多个组件通常拥有各种账号的配置信息，Service Account 是 Namespace 隔离的，可以针对组件进行一对一的定义，同时具备很好的“便携性”。

接下来，我们深入分析 Service Account 与 Secret 相关的一些运行机制。

前面的 Controller Manager 原理分析一节中，我们知道 Controller manager 创建了 ServiceAccount Controller 与 Token Controller 两个安全相关的控制器。其中 ServiceAccount Controller 一直监听 Service Account 和 Namespace 的事件，如果一个 Namespace 中没有 default Service Account，那么 ServiceAccount Controller 就会为该 Namespace 创建一个默认（default）的 Service Account，这就是我们之前看到每个 Namespace 下都有一个名为 default 的 Service Account 的原因了。

如果 Controller manager 进程在启动时指定了 API Server 私钥（service-account-private-key-file 参数），那么 Controller manager 会创建 Token Controller。Token Controller 也监听 Service Account 的事件，如果发现新创建的 Service Account 里没有对应的 Service Account Secret，则会用 API Server 私钥创建一个 Token（JWT Token），并用该 Token、CA 证书及 Namespace 名称等三个信息产生一个新的 Secret 对象，然后放入刚才的 Service Account 中；如果监听到事件是删除 Service Account 事件，则自动删除与该 Service Account 相关的所有 Secret。此外，Token Controller 对象同时监听 Secret 的创建、修改和删除事件，并根据事件的不同做不同的处理。

当我们在 API Server 的鉴权过程中启用了 Service Account 类型的准入控制器，即在 kube-apiserver 启动参数中包括下面的内容时：

```
--admission_control=ServiceAccount
```

则针对 Pod 新增或修改的请求，Service Account 准入控制器会验证 Pod 里的 Service Account 是否合法。

（1）如果 spec.serviceAccount 域没有被设置，则 Kubernetes 默认为其指定名字为 default 的 Service account。

（2）如果 Pod 的 spec.serviceAccount 域指定了 default 以外的 Service Account，而该 Service Account 没有事先被创建，则该 Pod 操作失败。

（3）如果在 Pod 中没有指定 “ImagePullSecrets”，那么这个 spec.serviceAccount 域指定的 Service Account 的 “ImagePullSecrets” 会被加入该 Pod 中。

（4）给 Pod 添加一个特殊的 Volume，在该 Volume 中包含 Service Account Secret 中的 Token，并将 Volume 挂载到 Pod 中所有容器的指定目录下（/var/run/secrets/kubernetes.io/serviceaccount）。

综上所述，Service Account 的正常工作离不开以下几个控制器。

（1）Admission Controller。



(2) Token Controller。

(3) ServiceAccount Controller。

### 3.6.5 Secret 私密凭据

上一节我们提到 Secret 对象, Secret 的主要作用是保管私密数据, 比如密码、OAuth Tokens、SSH Keys 等信息。将这些私密信息放在 Secret 对象中比直接放在 Pod 或 Docker Image 中更安全, 也更便于使用和分发。

下面的例子用于创建一个 Secret:

```
secrets.yaml:
apiVersion: v1
kind: Secret
metadata:
  name: mysecret
type: Opaque
data:
  password: dmFsdWUtMg0K
  username: dmFsdWUtMQ0K

# kubectl create -f secrets.yaml
```

在上面的例子中, data 域的各子域的值必须为 BASE64 编码值, 其中 password 域和 username 域 BASE64 编码前的值分别为 “value-1” 和 “value-2”。

一旦 Secret 被创建, 则可以通过下面的三种方式使用它。

(1) 在创建 Pod 时, 通过为 Pod 指定 Service Account 来自动使用该 Secret。

(2) 通过挂载该 Secret 到 Pod 来使用它。

(3) Docker 镜像下载时使用, 通过指定 Pod 的 spec.ImagePullSecrets 来引用它。

第 1 种使用方式主要用在 API Server 鉴权方面, 之前我们提到过。下面的例子展示了第 2 种使用方式: 将一个 Secret 通过挂载的方式添加到 Pod 的 Volume 中。

```
apiVersion: v1
kind: Pod
metadata:
  name: mypod
  namespace: myns
spec:
  containers:
  - name: mycontainer
    image: redis
```

```
volumeMounts:
- name: foo
  mountPath: "/etc/foo"
  readOnly: true
volumes:
- name: foo
  secret:
    secretName: mysecret
```

其结果如图 3.14 所示。

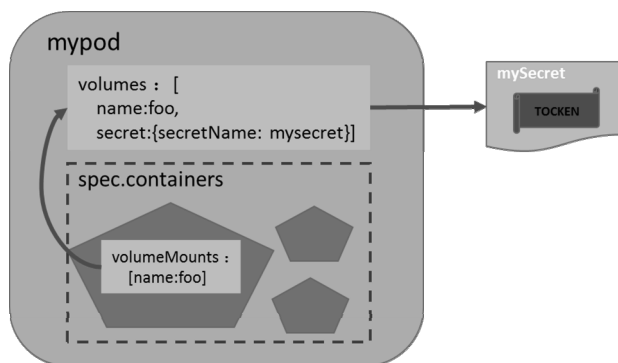


图 3.14 挂载 Secret 到 Pod

第 3 种使用方式的使用流程如下。

(1) 执行 login 命令，登录私有 Registry:

```
# docker login localhost:5000
```

输入用户名和密码，如果是第 1 次登录系统，则会创建新用户，相关信息会写入 ~/.dockercfg 文件中。

(2) 用 BASE64 编码 dockercfg 的内容:

```
# cat ~/.dockercfg | base64
```

(3) 将上一步命令的输出结果作为 Secret 的 “data.dockercfg” 域的内容，由此来创建一个 Secret:

```
image-pull-secret.yaml:
apiVersion: v1
kind: Secret
metadata:
  name: myregistrykey
data:
  .dockercfg: eyAiaHR0cHM6Ly9pbmRleC5kb2NrZXIuaW8vdjEvIjogeyAiYXV0aCI6ICJab
UZyWlhCaGMzTjNiM0prTVRJJSyIsICJlbWFPbCI6ICJqZG9lQG9lYUw1wGUuY29tIiB9IHOK
  type: kubernetes.io/dockercfg
```

```
# kubectl create -f image-pull-secret.yaml
```

(4) 在创建 Pod 时引用该 Secret:

```
pods.yaml:
apiVersion: v1
kind: Pod
metadata:
  name: mypod2
spec:
  containers:
    - name: foo
      image: janedoe/awesomeapp:v1
  imagePullSecrets:
    - name: myregistrykey
```

```
$ kubectl create -f pods.yaml
```

其结果如图 3.15 所示。

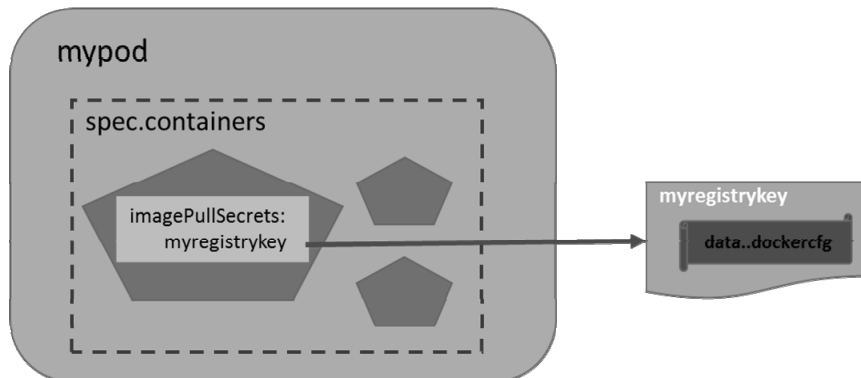


图 3.15 imagePullSecret 引用 Secret

每个单独的 Secret 大小不能超过 1MB，Kubernetes 不鼓励创建大尺寸的 Secret，因为如果使用大尺寸的 Secret，则将大量占用 API Server 和 kubelet 的内存。当然，创建许多小的 Secret 也能耗尽 API Server 和 kubelet 的内存。

在使用 Mount 方式挂载 Secret 时，Container 中 Secret 的“data”域的各个域的 Key 值作为目录中的文件，Value 值被 BASE64 编码后存储在相应的文件中。前面的例子中创建的 Secret，被挂载到一个叫作 mycontainer 的 Container 中，在该 Container 中可通过相应的查询命令查看所生成的文件和文件中的内容，如下所示：

```
$ ls /etc/foo/
username
password
```

```
$ cat /etc/foo/username
value-1
$ cat /etc/foo/password
value-2
```

通过上面的例子可以得出如下结论：我们可以通过 Secret 保管其他系统的敏感信息（比如数据库的用户名和密码），并以 Mount 的方式将 Secret 挂载到 Container 中，然后通过访问目录中的文件的方式获取该敏感信息。当 Pod 被 API Server 创建时，API Server 不会校验该 Pod 引用的 Secret 是否存在。一旦这个 Pod 被调度，则 kubelet 将试着获取 Secret 的值。如果 Secret 不存在或暂时无法连接到 API Server，则 kubelet 将按一定的时间间隔定期重试获取该 Secret，并发送一个 Event 来解释 Pod 没有启动的原因。一旦 Secret 被 Pod 获取，则 kubelet 将创建并 Mount 包含 Secret 的 Volume。只有所有 Volume 被 Mount 后，Pod 中的 Container 才会被启动。在 kubelet 启动 Pod 中的 Container 后，Container 中的和 Secret 相关的 Volume 将不会被改变，即使 Secret 本身被修改了。为了使用更新后的 Secret，必须删除旧的 Pod，并重新创建一个新的 Pod。

### 3.7 网络原理

关于 Kubernetes 网络，我们通常有这些问题需要回答，如图 3.16 所示。

Kubernetes 的网络模型是什么
Docker 背后的网络基础是什么
Docker 自身的网络模型和局限
Kubernetes 的网络组件之间是怎么通信的
外部如何访问 Kubernetes 的集群
有哪些开源的组件支持 Kubernetes 的网络模型

图 3.16 关于 Kubernetes 网络的常见问题

本节分别回答这些问题，然后通过一个具体的实验将这些相关的知识串联成一个整体。

#### 3.7.1 Kubernetes 网络模型

Kubernetes 网络模型设计的一个基础原则是：每个 Pod 都拥有一个独立的 IP 地址，而且假定所有 Pod 都在一个可以直接连通的、扁平的网络空间中。所以不管它们是否运行在同一个 Node（宿主机）中，都要求它们可以直接通过对方的 IP 进行访问。设计这个原则的原因是，用户不需要额外考虑如何建立 Pod 之间的连接，也不需要考虑将容器端口映射到主机端口等问题。

实际上在 Kubernetes 的世界里，IP 是以 Pod 为单位进行分配的。一个 Pod 内部的所有容器共享一个网络堆栈（实际上就是一个网络命名空间，包括它们的 IP 地址、网络设备、配置等都是共享的）。按照这个网络原则抽象出来的一个 Pod 一个 IP 的设计模型也被称作 IP-per-Pod 模型。

由于 Kubernetes 的网络模型假设 Pod 之间访问时使用的是对方 Pod 的实际地址，所以一个 Pod 内部的应用程序看到的自己的 IP 地址和端口与集群内其他 Pod 看到的一样。它们都是 Pod 实际分配的 IP 地址（从 docker0 上分配的）。将 IP 地址和端口在 Pod 内部和外部都保持一致，我们可以不使用 NAT 来进行转换，地址空间也自然是平的。Kubernetes 的网络之所以这么设计，主要原因就是可以兼容过去的应用。当然，我们使用 Linux 命令“ip addr show”也能看到这些地址，和程序看到的没有什么区别。所以这种 IP-per-Pod 的方案很好地利用了现有的各种域名解析和发现机制。

一个 Pod 一个 IP 的模型还有另外一层含义，那就是同一个 Pod 内的不同容器将会共享一个网络命名空间，也就是说同一个 Linux 网络协议栈。这就意味着同一个 Pod 内的容器可以通过 localhost 来连接对方的端口。这种关系和同一个 VM 内的进程之间的关系是一样的，看起来 Pod 内的容器之间的隔离性降低了，而且 Pod 内不同容器之间的端口是共享的，没有所谓的私有端口的概念了。如果你的应用必须要使用一些特定的端口范围，那么你也可以为这些应用单独创建一些 Pod。反之，对那些没有特殊需要的应用，这样做的好处是 Pod 内的容器是共享部分资源的，通过共享资源互相通信显然更加容易和高效。针对这些应用，虽然损失了可接受范围内的部分隔离性，但也是值得的。

IP-per-Pod 模式和 Docker 原生的通过动态端口映射方式实现的多节点访问模式有什么区别呢？主要区别是后者的动态端口映射会引入端口管理的复杂性，而且访问者看到的 IP 地址和端口与服务提供者实际绑定的不同（因为 NAT 的缘故，它们都被映射成新的地址或端口了），这也会引起应用配置的复杂化。同时，标准的 DNS 等名字解析服务也不适用了。甚至服务注册和发现机制都将受到挑战，因为在端口映射情况下，服务自身很难知道自己对外暴露的真实的服务 IP 和端口。而外部应用也无法通过服务所在容器的私有 IP 地址和端口来访问服务。

总的来说，IP-per-Pod 模型是一个简单的兼容性较好的模型。从该模型的网络的端口分配、域名解析、服务发现、负载均衡、应用配置和迁移等角度来看，Pod 都能够被看作一台独立的“虚拟机”或“物理机”。

按照这个网络抽象原则，Kubernetes 对网络有什么前提和要求呢？

Kubernetes 对集群的网络有如下要求。

- (1) 所有容器都可以在不用 NAT 的方式下同别的容器通信。
- (2) 所有节点都可以在不用 NAT 的方式下同所有容器通信，反之亦然。

（3）容器的地址和别人看到的地址是同一个地址。

这些基本的要求意味着并不是只要两台机器运行 Docker，Kubernetes 就可以工作了。具体的集群网络实现必须保障上述基本要求，原生的 Docker 网络目前还不能很好地支持这些要求。

实际上，这些对网络模型的要求并没有降低整个网络系统的复杂度。如果你的程序原来在 VM 上运行，而那些 VM 拥有独立 IP，并且它们之间可以直接透明地通信，那么 Kubernetes 的网络模型就和 VM 使用的网络模型是一样的。所以使用这种模型可以很容易地将已有的应用程序从 VM 或者物理机迁移到容器上。

当然，谷歌设计 Kubernetes 的一个主要运行基础就是其云环境 GCE（Google Compute Engine），在 GCE 下这些网络要求都是默认支持的。另外，常见的其他公有云服务商如亚马逊等，在它们的公有云计算环境下也是默认支持这个模型的。

由于部署私有云的场景会更普遍，所以在私有云中运行 Kubernetes+Docker 集群之前，就需要自己搭建出符合 Kubernetes 要求的网络环境。现在的开源世界有很多开源组件可以帮助我们打通 Docker 容器和容器之间的网络，实现 Kubernetes 要求的网络模型。当然每种方案都有适合的场景，我们要根据自己的实际需要进行选择。在后面的章节中会对常见的开源方案进行介绍。

Kubernetes 的网络依赖于 Docker，Docker 的网络又离不开 Linux 操作系统内核特性的支持，所以我们有必要先深入了解 Docker 背后的网络原理和基础知识。接下来我们一起深入学习一些必要的 Linux 网络知识。

### 3.7.2 Docker 的网络基础

Docker 本身的技术依赖于近年 Linux 内核虚拟化技术的发展，所以 Docker 对 Linux 内核的特性有很强的依赖。这里将 Docker 使用到的与 Linux 网络有关的主要技术进行简要介绍，这些技术包括如下几种，如图 3.17 所示。



图 3.17 Docker 使用到的与 Linux 网络有关的主要技术

## 1. 网络的命名空间

为了支持网络协议栈的多个实例，Linux 在网络栈中引入了网络命名空间（Network Namespace），这些独立的协议栈被隔离到不同的命名空间中。处于不同命名空间的网络栈是完全隔离的，彼此之间无法通信，就好像两个“平行宇宙”。通过这种对网络资源的隔离，就能在一个宿主机上虚拟多个不同的网络环境。而 Docker 也正是利用了网络的命名空间特性，实现了不同容器之间网络的隔离。

在 Linux 的网络命名空间内可以有自己独立的路由表及独立的 Iptables/Netfilter 设置来提供包转发、NAT 及 IP 包过滤等功能。

为了隔离出独立的协议栈，需要纳入命名空间的元素有进程、套接字、网络设备等。进程创建的套接字必须属于某个命名空间，套接字的操作也必须在命名空间内进行。同样，网络设备也必须属于某个命名空间。因为网络设备属于公共资源，所以可以通过修改属性实现在命名空间之间移动。当然，是否允许移动和设备的特征有关。

让我们稍微深入 Linux 操作系统内部，看它是如何实现网络命名空间的，这也会对理解后面的概念有帮助。

### 1) 网络命名空间的实现

Linux 的网络协议栈是十分复杂的，为了支持独立的协议栈，相关的这些全局变量都必须修改为协议栈私有。最好的办法就是让这些全局变量成为一个 Net Namespace 变量的成员，然后为协议栈的函数调用加入一个 Namespace 参数。这就是 Linux 实现网络命名空间的核心。

同时，为了保证对已经开发的应用程序及内核代码的兼容性，内核代码隐式地使用了命名空间内的变量。我们的程序如果没有对命名空间的特殊需求，那么不需要写额外的代码，网络命名空间对应用程序而言是透明的。

在建立了新的网络命名空间，并将某个进程关联到这个网络命名空间后，就出现了类似于如图 3.18 所示的内核数据结构，所有网络栈变量都放入了网络命名空间的数据结构中。这个网络命名空间是属于它的进程组私有的，和其他进程组不冲突。

新生成的私有命名空间只有回环 lo 设备（而且是停止状态），其他设备默认都不存在，如果我们需要，则要一一手工建立。Docker 容器中的各类网络栈设备都是 Docker Daemon 在启动时自动创建和配置的。

所有的网络设备（物理的或虚拟接口、桥等在内核里都叫作 Net Device）都只能属于一个命名空间。当然，通常物理的设备（连接实际硬件的设备）只能关联到 root 这个命名空间中。虚拟的网络设备（虚拟的以太网接口或者虚拟网口对）则可以被创建并关联到一个给定的命名空间中，而且可以在这些命名空间之间移动。

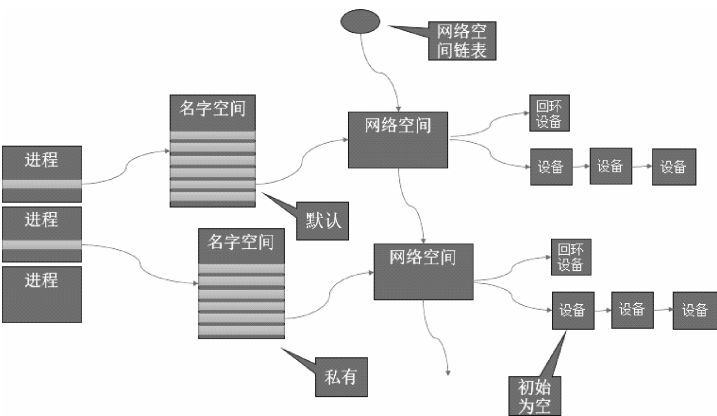


图 3.18 命名空间的内核数据结构

前面我们提到，由于网络命名空间代表的是一个独立的协议栈，所以它们之间是相互隔离的，彼此无法通信，在协议栈内部都看不到对方。那么有没有办法打破这种限制，让处于不同命名空间的网络相互通信，甚至和外部的网络进行通信呢？答案就是“Veth 设备对”。“Veth 设备对”的一个重要作用就是打通互相看不到的协议栈之间的壁垒，它就像一个管子，一端连着这个网络命名空间的协议栈，一端连着另一个网络命名空间的协议栈。所以如果想在两个命名空间之间进行通信，就必须有一个 Veth 设备对。后面我们会介绍如何操作 Veth 设备对来打通不同命名空间之间的网络。

2) 网络命名空间操作

下面列举一些网络命名空间的操作。

我们可以使用 Linux iproute2 系列配置工具中的 IP 命令来操作网络命名空间。注意，这个命令需要由 root 用户运行。

创建一个命名空间：

```
ip netns add <name>
```

在命名空间内执行命令：

```
ip netns exec <name> <command>
```

如果想执行多个命令，则可以先进入内部的 sh，然后执行：

```
ip netns exec <name> bash
```

之后就是新的命名空间内进行操作了。退出到外面的命名空间时，请输入“exit”。

3) 网络命名空间的一些技巧

操作网络命名空间时的一些实用技巧如下。



我们可以在不同的网络命名空间之间转移设备，例如下面会提到的 Veth 设备对的转移。因为一个设备只能属于一个命名空间，所以转移后在这个命名空间内就看不到这个设备了。具体哪些设备能够转移到不同的命名空间呢？在设备里面有一个重要的属性：NETIF\_F\_ETNS\_LOCAL，如果这个属性为“on”，则不能转移到其他命名空间内。Veth 设备属于可以转移的设备，而很多其他设备如 lo 设备、vxlan 设备、ppp 设备、bridge 设备等都是不可以转移的。至于将无法转移的设备移动到别的命名空间的操作，则会得到无效参数的错误提示。

```
# ip link set br0 netns ns1
RTNETLINK answers: Invalid argument
```

如何知道这些设备是否可以转移呢？可以使用 ethtool 工具查看：

```
# ethtool -k br0
netns-local: on [fixed]
```

netns-local 的值是 on，就说明不可以转移，否则可以。

## 2. Veth 设备对

引入 Veth 设备对是为了在不同的网络命名空间之间进行通信，利用它可以直接将两个网络命名空间连接起来。由于要连接两个网络命名空间，所以 Veth 设备都是成对出现的，很像一对以太网卡，并且中间有一根直连的网线。既然是一对网卡，那么我们将其中一端称为另一端的 peer。在 Veth 设备的一端发送数据时，它会将数据直接发送到另一端，并触发另一端的接收操作。

整个 Veth 的实现非常简单，有兴趣的读者可以参考源代码“drivers/net/veth.c”的实现。图 3.19 是 Veth 设备对的示意图。

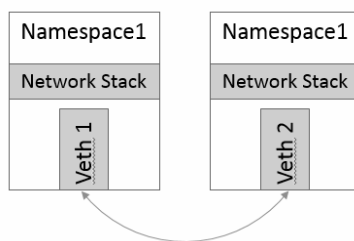


图 3.19 Veth 设备对示意图

### 1) Veth 设备对的操作命令

接下来看看如何创建 Veth 设备对，如何连接到不同的命名空间，并设置它们的地址，让它们通信。

创建 Veth 设备对：

```
ip link add veth0 type veth peer name veth1
```

创建后，可以查看 Veth 设备对的信息。使用 `ip link show` 命令查看所有网络接口：

```
# ip link show
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
    Link/loopback: 00:00:00:00:00:00 brd 00:00:00:00:00:00
2: eno16777736: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
state UP mode DEFAULT qlen 1000
    link/ether 00:0c:29:cf:1a:2e brd ff:ff:ff:ff:ff:ff
3: docker0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc noqueue state UP
mode DEFAULT
    link/ether 56:84:7a:fe:97:99 brd ff:ff:ff:ff:ff:ff
19: veth1: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN mode DEFAULT qlen 1000
    link/ether 7e:4a:ae:41:a3:65 brd ff:ff:ff:ff:ff:ff
20: veth0: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN mode DEFAULT qlen 1000
    link/ether ea:da:85:a3:75:8a brd ff:ff:ff:ff:ff:ff
```

看到了吧，有两个设备生成了，一个是 `veth0`，它的 `peer` 是 `veth1`。

现在这两个设备都在自己的命名空间内，那怎么能行呢？好了，如果将 Veth 看作有两个头的网线，那么我们将另一个头甩给另一个命名空间吧：

```
ip link set veth1 netns netns1
```

这时可在外面这个命名空间内看两个设备的情况：

```
# ip link show
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
    Link/loopback: 00:00:00:00:00:00 brd 00:00:00:00:00:00
2: eno16777736: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
state UP mode DEFAULT qlen 1000
    link/ether 00:0c:29:cf:1a:2e brd ff:ff:ff:ff:ff:ff
3: docker0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc noqueue state UP
mode DEFAULT
    link/ether 56:84:7a:fe:97:99 brd ff:ff:ff:ff:ff:ff
20: veth0: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN mode DEFAULT qlen 1000
    link/ether ea:da:85:a3:75:8a brd ff:ff:ff:ff:ff:ff
```

只剩一个 `veth0` 设备了，已经看不到另一个设备了，另一个设备已经转移到另一个网络命名空间了。

在 `netns1` 网络命名空间中可以看到 `veth1` 设备了，符合预期。

```
# ip netns exec netns1 ip link show
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN mode DEFAULT
    Link/loopback: 00:00:00:00:00:00 brd 00:00:00:00:00:00
19: veth1: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN mode DEFAULT qlen 1000
    link/ether 7e:4a:ae:41:a3:65 brd ff:ff:ff:ff:ff:ff
```

现在看到的结果是，两个不同的命名空间各自有一个 Veth 的“网线头”，各显示为一个 Device

(在 Docker 的实现里面, 它除了将 Veth 放入容器内, 还将它的名字改成了 eth0, 简直以假乱真, 你以为它是一个本地网卡吗)。

现在可以通信了吗? 不行, 因为它们还没有任何地址, 现在我们来给它们分配 IP 地址吧:

```
ip netns exec netns1 ip addr add 10.1.1.1/24 dev veth1
ip addr add 10.1.1.2/24 dev veth0
```

再启动它们:

```
ip netns exec netns1 ip link set dev veth1 up
ip link set dev veth0 up
```

现在两个网络命名空间可以互相通信了:

```
# ping 10.1.1.1
PING 10.1.1.1 (10.1.1.1) 56(84) bytes of data.
64 bytes from 10.1.1.1: icmp_seq=1 ttl=64 time=0.035 ms
64 bytes from 10.1.1.1: icmp_seq=2 ttl=64 time=0.096 ms
^C
--- 10.1.1.1 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1001ms
rtt min/avg/max/mdev = 0.035/0.065/0.096/0.031 ms

# ip netns exec netns1 ping 10.1.1.2
PING 10.1.1.2 (10.1.1.2) 56(84) bytes of data.
64 bytes from 10.1.1.2: icmp_seq=1 ttl=64 time=0.045 ms
64 bytes from 10.1.1.2: icmp_seq=2 ttl=64 time=0.105 ms
^C
--- 10.1.1.2 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1000ms
rtt min/avg/max/mdev = 0.045/0.075/0.105/0.030 ms
```

至此, 两个网络命名空间之间就完全通了。

至此我们就能理解 Veth 设备对的原理和用法了。在 Docker 内部, Veth 设备对也是联系容器到外面的重要设备, 离开它是不行的。

## 2) Veth 设备对如何查看对端

我们在操作 Veth 设备对时有一些实用技巧, 如下所示。

一旦将 Veth 设备对的 peer 端放入另一个命名空间, 我们在本命名空间内就看不到它了。那么我们怎么知道这个 Veth 对的对端在哪里呢, 也就是说它到底连接到哪个别的命名空间呢? 可以使用 ethtool 工具来查看 (当网络命名空间特别多时, 这可不是一件很容易的事情)。

首先我们在一个命名空间中查询 Veth 设备对端接口在设备列表中的序列号:

```
ip netns exec netns1 ethtool -S veth1
NIC statistics:
    peer_ifindex: 5
```

得知另一端的接口设备的序列号是 5，我们再到另一个命名空间中查看序列号 5 代表什么设备：

```
ip netns exec netns2 ip link | grep 5    <-- 我们只关注序列号是 5 的设备
veth0
```

好了，我们现在就找到下标为 5 的设备了，它是 veth0，它的另一端自然就是另一个命名空间中的 veth1 了，因为它们互为 peer。

### 3. 网桥

Linux 可以支持多个不同的网络，它们之间能够相互通信，如何将这些网络连接起来并实现各网络中主机的相互通信呢？这就是网桥的作用了。网桥是一个二层的虚拟网络设备，把若干个网络接口“连接”起来，以使得网口之间的报文能够互相转发。网桥能够解析收发的报文，读取目标 MAC 地址的信息，和自己记录的 MAC 表结合，来决策报文的转发目标网口。为了实现这些功能，网桥会学习源 MAC 地址（二层网桥转发的依据就是 MAC 地址）。在转发报文时，网桥只需要向特定的网口进行转发，从而避免不必要的网络交互。如果它遇到一个自己从未学习到的地址，就无法知道这个报文应该向哪个网口转发，就将报文广播给所有的网口（报文来源的网口除外）。

在实际网络中，网络拓扑不可能永久不变。如果设备移动到另一个端口上，而它没有发送任何数据，那么网桥设备就无法感知到这个变化，结果网桥还是向原来的端口转发数据包，在这种情况下数据就会丢失。所以网桥还要对学习到的 MAC 地址表加上超时时间（默认为 5min）。如果网桥收到了对应端口 MAC 地址回发的包，则重置超时时间，否则过了超时时间后，就认为那个设备已经不在那个端口上了，它就会重新广播发送。

在 Linux 的内部网络栈里面实现的网桥设备，作用和上面的描述相同。过去 Linux 主机一般都只有一个网卡，现在多网卡的机器越来越多，而且还有很多虚拟的设备存在，所以 Linux 的网桥提供了这些设备之间互相转发数据的二层设备。

Linux 内核支持网口的桥接（目前只支持以太网接口）。但是与单纯的交换机不同，交换机只是一个二层设备，对于接收到的报文，要么转发，要么丢弃。运行着 Linux 内核的机器本身就是一台主机，有可能是网络报文的目的地，其收到的报文除了转发和丢弃，还可能被送到网络协议栈的上层（网络层），从而被自己（这台主机本身的协议栈）消化，所以我们既可以把网桥看作一个二层设备，也可以看作一个三层设备。

#### 1) Linux 网桥的实现

Linux 内核是通过一个虚拟的网桥设备（Net Device）来实现桥接的。这个虚拟设备可以绑定若干个以太网接口设备，从而将它们桥接起来。如图 3.20 所示，这种 Net Device 网桥和普通

的设备不同，最明显的一个特性是它还可以有一个 IP 地址。

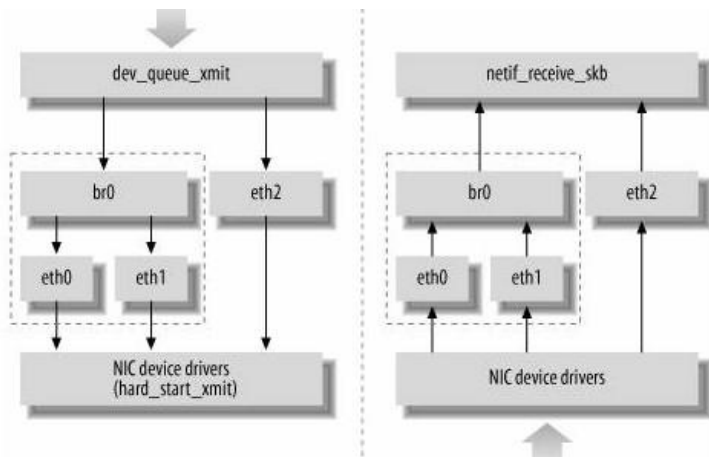


图 3.20 网桥的位置

如图 3.20 所示，网桥设备 br0 绑定了 eth0 和 eth1。对于网络协议栈的上层来说，只看得到 br0。因为桥接是在数据链路层实现的，上层不需要关心桥接的细节，于是协议栈上层需要发送的报文被送到 br0，网桥设备的处理代码判断报文该被转发到 eth0 还是 eth1，或者两者皆转发；反过来，从 eth0 或从 eth1 接收到的报文被提交给网桥的处理代码，在这里会判断报文应该被转发、丢弃还是提交到协议栈上层。

而有时 eth0、eth1 也可能会作为报文的源地址或目的地址，直接参与报文的发送与接收，从而绕过网桥。

## 2) 网桥的常用操作命令

Docker 自动完成了对网桥的创建和维护。为了进一步理解网桥，下面举几个常用的网桥操作例子，对网桥进行手工操作：

新增一个网桥设备：

```
#brctl addbr xxxxx
```

之后可以为网桥增加网口，在 Linux 中，一个网口其实就是一个物理网卡。将物理网卡和网桥连接起来：

```
#brctl addif xxxxx ethx
```

网桥的物理网卡作为一个网口，由于在链路层工作，就不再需要 IP 地址了，这样上面的 IP 地址自然失效：

```
#ifconfig ethx 0.0.0.0
```

给网桥配置一个 IP 地址：

```
#ifconfig brxxx xxx.xxx.xxx.xxx
```

这样网桥就有了一个 IP 地址，而连接到上面的网卡就是一个纯链路层设备了。

4. iptables/Netfilter

我们知道，Linux 网络协议栈非常高效，同时比较复杂。如果我们希望在数据的处理过程中对关心的数据进行一些操作该怎么做呢？Linux 提供了一套机制来为用户实现自定义的数据包处理过程。

在 Linux 网络协议栈中有一组回调函数挂接点，通过这些挂接点挂接的钩子函数可以在 Linux 网络栈处理数据包的过程中对数据包进行一些操作，例如过滤、修改、丢弃等。整个挂接点技术叫作 Netfilter 和 Iptables。

Netfilter 负责在内核中执行各种挂接的规则，运行在内核模式中；而 Iptables 是在用户模式下运行的进程，负责协助维护内核中 Netfilter 的各种规则表。通过二者的配合来实现整个 Linux 网络协议栈中灵活的数据包处理机制。

Netfilter 可以挂接的规则点有 5 个，如图 3.21 中的深色椭圆所示。

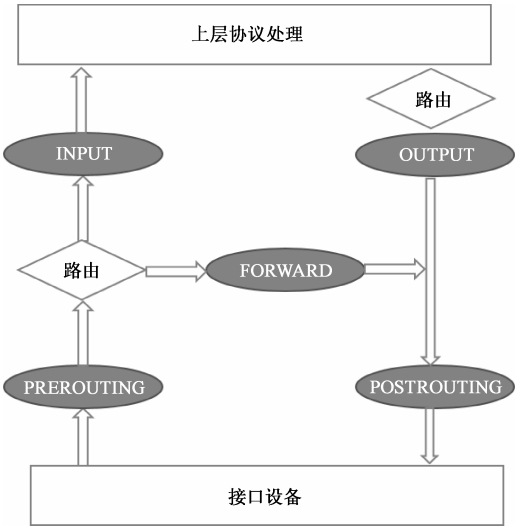


图 3.21 Netfilter 可以挂接的规则点

1) 规则表 Table

这些挂接点能挂接的规则也分不同的类型（也就是规则表 Table），我们可以在不同类型的 Table 中加入我们的规则。目前主要支持的 Table 类型如下。

- ◎ RAW。
- ◎ MANGLE。
- ◎ NAT。
- ◎ FILTER。

上述4个Table（规则链）的优先级是RAW最高，FILTER最低。

在实际应用中，不同的挂接点需要的规则类型通常不同。例如，在Input的挂接点上明显不需要FILTER过滤规则，因为根据目标地址，已经选择好本机的上层协议栈了，所以无须再挂接FILTER过滤规则。目前Linux系统支持的不同挂接点能挂接的规则类型如图3.22所示。

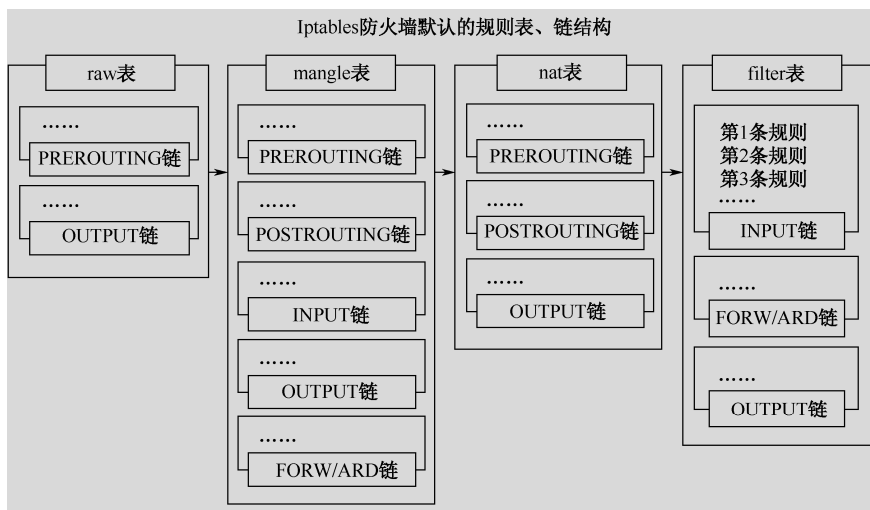


图 3.22 不同挂接点能挂接的规则类型

当Linux协议栈的数据处理运行到挂接点时，它会依次调用挂接点上所有的挂钩函数，直到数据包的处理结果是明确地接受或者拒绝。

## 2) 处理规则

每个规则的特性都分为以下几部分。

- ◎ 表类型（准备干什么事情）。
- ◎ 什么挂接点（什么时候起作用）。
- ◎ 匹配的参数是什么（针对什么样的数据包）。
- ◎ 匹配后有什么动作（匹配后具体的操作是什么）。

表类型和什么挂接点在前面已经介绍了，现在我们看看匹配的参数和匹配后的动作。

### （1）匹配的参数

匹配的参数用于对数据包或者 TCP 数据连接的状态进行匹配。当有多个条件存在时，它们一起起作用，来达到只针对某部分数据进行修改的目的。常见的匹配参数如下。

- ◎ 流入、流出的网络接口。
- ◎ 来源、目的地址。
- ◎ 协议类型。
- ◎ 来源、目的端口。

### （2）匹配后的动作

一旦有数据匹配上，就会执行相应的动作。动作类型既可以是标准的预定义的几个动作，也可以是自定义的模块注册的动作，或者是一个新的规则链，以便更好地组织一组动作。

## 3) Iptables 命令

Iptables 命令用于协助用户维护各种规则。我们在使用 Kubernetes、Docker 的过程中，通常都会去查看相关的 Netfilter 配置。这里只介绍如何查看规则表，详细的介绍请参照 Linux 的 Iptables 帮助文档。

查看系统中已有的规则的方法如下。

- ◎ iptables-save：按照命令的方式打印 Iptables 的内容。
- ◎ Iptables-vnL：以另一种格式显示 Netfilter 表的内容。

## 5. 路由

Linux 系统包含一个完整的路由功能。当 IP 层在处理数据发送或者转发时，会使用路由表来决定发往哪里。通常情况下，如果主机与目的主机直接相连，那么主机可以直接发送 IP 报文到目的主机，这个过程比较简单。例如，通过点对点的链接或通过网络共享，如果主机与目的主机没有直接相连，那么主机会将 IP 报文发送给默认的路由器，然后由路由器来决定往哪发送 IP 报文。

路由功能由 IP 层维护的一张路由表来实现。当主机收到数据报文时，它用此表来决策接下来应该做什么操作。当从网络侧接收到数据报文时，IP 层首先会检查报文的 IP 地址是否与主机自身的地址相同。如果数据报文中的 IP 地址是主机自身的地址，那么报文将被发送到传输层相应的协议中去。如果报文中的 IP 地址不是主机自身的地址，并且主机配置了路由功能，那么报文将被转发，否则，报文将被丢弃。



路由表中的数据一般是以条目形式存在的。一个典型的路由表条目通常包含以下主要的条目项。

(1) 目的 IP 地址：此字段表示目标的 IP 地址。这个 IP 地址可以是某台主机的地址，也可以是一个网络地址。如果这个条目包含的是一个主机地址，那么它的主机 ID 将被标记为非零；如果这个条目包含的是一个网络地址，那么它的主机 ID 将被标记为零。

(2) 下一个路由器的 IP 地址：为什么采用“下一个”的说法，是因为下一个路由器并不总是最终的目的路由器，它很可能是一个中间路由器。条目给出下一个路由器的地址用来转发从相应接口接收到的 IP 数据报文。

(3) 标志：这个字段提供了另一组重要信息，例如目的 IP 地址是一个主机地址还是一个网络地址。此外，从标志中可以得知下一个路由器是一个真实路由器还是一个直接相连的接口。

(4) 网络接口规范：为一些数据报文的网络接口规范，该规范将与报文一起被转发。

在通过路由表转发时，如果任何条目的第 1 个字段完全匹配目的 IP 地址（主机）或部分匹配条目的 IP 地址（网络），那么它将指示下一个路由器的 IP 地址。这是一个重要的信息，因为这些信息直接告诉主机（具备路由功能的）数据包应该转发到哪个“下一个路由器”去。而条目中的所有其他字段将提供更多的辅助信息来为路由转发做决定。

如果没有找到一个完全匹配的 IP，那么就接着搜索相匹配的网络 ID。如果找到，那么该数据报文会被转发到指定的路由器上。可以看出，网络上的所有主机都通过这个路由表中的单个（这个）条目进行管理。

如果上述两个条件都不匹配，那么该数据报文将被转发到一个默认路由器上。

如果上述步骤失败，默认路由器也不存在，那么该数据报文最终无法被转发。任何无法投递的数据报文都将产生一个 ICMP 主机不可达或 ICMP 网络不可达的错误，并将此错误返回给生成此数据报文的应用程序。

### 1) 路由表的创建

Linux 的路由表至少包括两个表（当启用策略路由时，还会有其他表）：一个是 LOCAL，另一个是 MAIN。在 LOCAL 表中会包含所有的本地设备地址。LOCAL 路由表是在配置网络设备地址时自动创建的。LOCAL 表用于供 Linux 协议栈识别本地地址，以及进行本地各个不同网络接口之间的数据转发。

可以通过下面的命令查看 LOCAL 表的内容：

```
# ip route show table local type local
10.1.1.0 dev flannel0 proto kernel scope host src 10.1.1.0
127.0.0.0/8 dev lo proto kernel scope host src 127.0.0.1
127.0.0.1 dev lo proto kernel scope host src 127.0.0.1
```

```
172.17.42.1 dev docker proto kernel scope host src 172.17.42.1
192.168.1.128 dev eno16777736 proto kernel scope host src 192.168.1.128
```

MAIN 表用于各类网络 IP 地址的转发。它的建立既可以使用静态配置生成，也可以使用动态路由发现协议生成。动态路由发现协议一般使用组播功能来通过发送路由发现数据，动态地交换和获取网络的路由信息，并更新到路由表中。

Linux 下支持路由发现协议的开源软件有许多，常用的有 Quagga、Zebra 等。第 4 章会介绍使用 Quagga 动态容器路由发现的机制来实现 Kubernetes 的网络组网。

## 2) 路由表的查看

我们可以使用 `ip route list` 命令查看当前的路由表。

```
# ip route list
192.168.6.0/24 dev eno16777736 proto kernel scope link src 192.168.6.140
metric 1
```

在上面的例子代码中，只有一个子网的路由，源地址是 192.168.6.140（本机），目标地址是 192.168.6.0/24 网段的数据，都将通过 `eth0` 接口设备发送出去。

Netstat-rn 是另一个查看路由表的工具：

```
# netstat -rn
Kernel IP routing table
Destination      Gateway          Genmask         Flags         MSS Window  irtt Iface
0.0.0.0          192.168.6.2     0.0.0.0         UG            0 0          0 eth0
192.168.6.0      0.0.0.0         255.255.255.0   U             0 0          0 eth0
```

在它显示的信息中，如果标志是 U，则说明是可达路由；如果标志是 G，则说明这个网络接口连接的是网关，否则说明是直连主机。

## 3.7.3 Docker 的网络实现

标准的 Docker 支持以下 4 类网络模式。

- ◎ host 模式：使用 `--net=host` 指定。
- ◎ container 模式：使用 `--net=container:NAME_or_ID` 指定。
- ◎ none 模式：使用 `--net=none` 指定。
- ◎ bridge 模式：使用 `--net=bridge` 指定，为默认设置。

在 Kubernetes 管理模式下，通常只会使用 bridge 模式，所以本节只介绍 bridge 模式下 Docker 是如何支持网络的。

在 bridge 模式下，Docker Daemon 第 1 次启动时会创建一个虚拟的网桥，默认的名字是

docker0, 然后按照 RPC1918 的模型, 在私有网络空间中给这个网桥分配一个子网。针对由 Docker 创建出来的每一个容器, 都会创建一个虚拟的以太网设备 (Veth 设备对), 其中一端关联到网桥上, 另一端使用 Linux 的网络命名空间技术, 映射到容器内的 eth0 设备, 然后从网桥的地址段内给 eth0 接口分配一个 IP 地址。

如图 3.23 所示就是 Docker 的默认桥接网络模型。

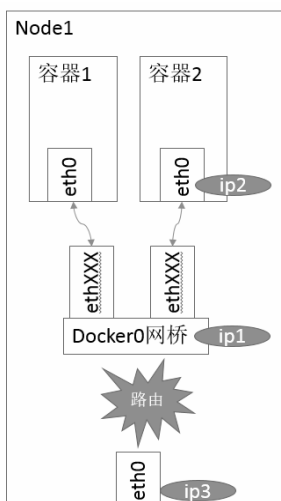


图 3.23 Docker 的默认桥接网络模型

其中 ip1 是网桥的 IP 地址, Docker Daemon 会在几个备选地址段里给它选一个, 通常是 172 开头的地址。这个地址和主机的 IP 地址是不重叠的。ip2 是 Docker 在启动容器时, 在这个地址段随机选择的一个没有使用的 IP 地址, Docker 占用它并分配给了被启动的容器。相应的 MAC 地址也根据这个 IP 地址, 在 02:42:ac:11:00:00 和 02:42:ac:11:ff:ff 的范围内生成, 这样做可以确保不会有 ARP 的冲突。

启动后, Docker 还将 Veth 对的名字映射到了 eth0 网络接口。ip3 就是主机的网卡地址。

在一般情况下, ip1、ip2 和 ip3 是不同的 IP 段, 所以在默认不做任何特殊配置的情况下, 在外部是看不到 ip1 和 ip2 的。

这样做的结果就是, 同一台机器内的容器之间可以相互通信。不同主机上的容器不能够相互通信。实际上它们甚至有可能在相同的网络地址范围内 (不同的主机上的 docker0 的地址段可能是一样的)。

为了让它们跨节点互相通信, 就必须在主机的地址上分配端口, 然后通过这个端口路由或代理到容器上。这种做法显然意味着一定要在容器之间小心谨慎地协调好端口的分配, 或者使

用动态端口的分配技术。在不同应用之间协调好端口分配是十分困难的事情，特别是集群水平扩展时。而动态的端口分配也会带来高度复杂性，例如：每个应用程序都只能将端口看作一个符号（因为是动态分配的，无法提前设置）。而且 API Server 也要在分配完后，将动态端口插入到配置的合适位置。另外，服务也必须能互相之间找到对方等。这些都是 Docker 的网络模型在跨主机访问时面临的问题。

### 1) 查看 Docker 启动后的系统情况

我们已经知道，Docker 网络在 bridge 模式下 Docker Daemon 启动时创建 docker0 网桥，并在网桥使用的网段为容器分配 IP。让我们看看实际的操作。

在刚刚启动 Docker Daemon 并且还没有启动任何容器时，网络协议栈的配置情况如下：

```
# systemctl start docker
# ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
        valid_lft forever preferred_lft forever
2: eno16777736: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
state UP qlen 1000
    link/ether 00:0c:29:14:3d:80 brd ff:ff:ff:ff:ff:ff
    inet 192.168.1.133/24 brd 192.168.1.255 scope global eno16777736
        valid_lft forever preferred_lft forever
    inet6 fe80::20c:29ff:fe14:3d80/64 scope link
        valid_lft forever preferred_lft forever
3: docker0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc noqueue state DOWN
    link/ether 02:42:6e:af:0e:c3 brd ff:ff:ff:ff:ff:ff
    inet 172.17.42.1/24 scope global docker0
        valid_lft forever preferred_lft forever

# iptables-save
# Generated by iptables-save v1.4.21 on Thu Sep 24 17:11:04 2015
*nat
:PREROUTING ACCEPT [7:878]
:INPUT ACCEPT [7:878]
:OUTPUT ACCEPT [3:536]
:POSTROUTING ACCEPT [3:536]
:DOCKER - [0:0]
-A PREROUTING -m addrtype --dst-type LOCAL -j DOCKER
-A OUTPUT ! -d 127.0.0.0/8 -m addrtype --dst-type LOCAL -j DOCKER
-A POSTROUTING -s 172.17.0.0/16 ! -o docker0 -j MASQUERADE
COMMIT
# Completed on Thu Sep 24 17:11:04 2015
```

```
# Generated by iptables-save v1.4.21 on Thu Sep 24 17:11:04 2015
*filter
:INPUT ACCEPT [133:11362]
:FORWARD ACCEPT [0:0]
:OUTPUT ACCEPT [37:5000]
:DOCKER - [0:0]
-A FORWARD -o docker0 -j DOCKER
-A FORWARD -o docker0 -m conntrack --ctstate RELATED,ESTABLISHED -j ACCEPT
-A FORWARD -i docker0 ! -o docker0 -j ACCEPT
-A FORWARD -i docker0 -o docker0 -j ACCEPT
COMMIT
# Completed on Thu Sep 24 17:11:04 2015
```

可以看到，Docker 创建了 docker0 网桥，并添加了 Iptables 规则。docker0 网桥和 Iptables 规则都处于 root 命名空间中。通过解读这些规则，我们发现，在还没有启动任何容器时，如果启动了 Docker Daemon，那么它就已经做好了通信的准备。对这些规则的说明如下。

(1) 在 NAT 表中有 3 条记录，前两条匹配生效后，都会继续执行 DOCKER 链，而此时 DOCKER 链为空，所以前两条只是做了个框架，并没有实际效果。

(2) NAT 表第 3 条的含义是，若本地发出的数据包不是发往 docker0 的，即是发往主机之外的设备的，都需要进行动态地址修改（MASQUERADE），将源地址从容器的地址（172 段）修改为宿主机网卡的 IP 地址，之后就可以发送给外面的网络了。

(3) 在 FILTER 表中，第 1 条也是一个框架，因为后继的 DOCKER 链是空的。

(4) 在 FILTER 表中，第 3 条是说，docker0 发出的包，如果需要 Forward 到非 docker0 的本地 IP 地址的设备，则是允许的，这样，docker0 设备的包就可以根据路由规则中转到宿主机的网卡设备，从而访问外面的网络。

(5) FILTER 表中，第 4 条是说，docker0 的包还可以中转给 docker0 本身，即连接在 docker0 网桥上的不同容器之间的通信也是允许的。

(6) FILTER 表中，第 2 条是说，如果接收到的数据包属于以前已经建立好的连接，那么允许直接通过。这样接收到的数据包自然又走向 docker0，并中转到相应的容器。

除了这些 Netfilter 的设置，Linux 的 ip\_forward 功能也被 Docker Daemon 打开了：

```
# cat /proc/sys/net/ipv4/ip_forward
1
```

另外，我们还可以看到刚刚启动 Docker 后的 Route 表，和启动前没有什么不同：

```
# ip route
default via 192.168.1.2 dev eno16777736 proto static metric 100
172.17.0.0/16 dev docker proto kernel scope link src 172.17.42.1
192.168.1.0/24 dev eno16777736 proto kernel scope link src 192.168.1.132
```

```
192.168.1.0/24 dev enol6777736 proto kernel scope link src 192.168.1.132
metric 100
```

## 2) 查看容器启动后的情况（容器无端口映射）

刚才我们看了 Docker 服务启动后的网络情况。现在，我们启动一个 Registry 容器后（不使用任何端口镜像参数），看一下网络堆栈部分相关的变化：

```
docker run --name register -d registry
# ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
        valid_lft forever preferred_lft forever
2: enol6777736: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
state UP qlen 1000
    link/ether 00:0c:29:c8:12:5f brd ff:ff:ff:ff:ff:ff
    inet 192.168.1.132/24 brd 192.168.1.255 scope global enol6777736
        valid_lft forever preferred_lft forever
    inet6 fe80::20c:29ff:fec8:125f/64 scope link
        valid_lft forever preferred_lft forever
3: docker0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc noqueue state DOWN
    link/ether 02:42:72:79:b8:88 brd ff:ff:ff:ff:ff:ff
    inet 172.17.42.1/24 scope global docker0
        valid_lft forever preferred_lft forever
    inet6 fe80::42:7aff:fe79:b888/64 scope link
        valid_lft forever preferred_lft forever
13: veth2dc8bbd: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue master
docker0 state UP
    link/ether be:d9:19:42:46:18 brd ff:ff:ff:ff:ff:ff
    inet6 fe80::bcd9:19ff:fe42:4618/64 scope link
        valid_lft forever preferred_lft forever

# iptables-save
# Generated by iptables-save v1.4.21 on Thu Sep 24 18:21:04 2015
*nat
:PREROUTING ACCEPT [14:1730]
:INPUT ACCEPT [14:1730]
:OUTPUT ACCEPT [59:4918]
:POSTROUTING ACCEPT [59:4918]
:DOCKER - [0:0]
-A PREROUTING -m addrtype --dst-type LOCAL -j DOCKER
-A OUTPUT ! -d 127.0.0.0/8 -m addrtype --dst-type LOCAL -j DOCKER
-A POSTROUTING -s 172.17.0.0/16 ! -o docker0 -j MASQUERADE
COMMIT
```

```
# Completed on Thu Sep 24 18:21:04 2015
# Generated by iptables-save v1.4.21 on Thu Sep 24 18:21:04 2015
*filter
:INPUT ACCEPT [2383:211572]
:FORWARD ACCEPT [0:0]
:OUTPUT ACCEPT [2004:242872]
:DOCKER - [0:0]
-A FORWARD -o docker0 -j DOCKER
-A FORWARD -o docker0 -m conntrack --ctstate RELATED,ESTABLISHED -j ACCEPT
-A FORWARD -i docker0 ! -o docker0 -j ACCEPT
-A FORWARD -i docker0 -o docker0 -j ACCEPT
COMMIT
# Completed on Thu Sep 24 18:21:04 2015

# ip route
default via 192.168.1.2 dev eno16777736 proto static metric 100
172.17.0.0/16 dev docker proto kernel scope link src 172.17.42.1
192.168.1.0/24 dev eno16777736 proto kernel scope link src 192.168.1.132
192.168.1.0/24 dev eno16777736 proto kernel scope link src 192.168.1.132
metric 100
```

可以看到如下情况。

(1) 宿主机上的 **Netfilter** 和路由表都没有变化，说明在不进行端口映射时，**Docker** 的默认网络是没有特殊处理的。相关的 **NAT** 和 **FILTER** 两个 **Netfilter** 链还是空的。

(2) 宿主机上的 **Veth** 对已经建立，并连接到了容器内。

我们再次进入刚刚启动的容器内，看看网络栈是什么情况。容器内部的 **IP** 地址和路由如下：

```
# docker exec -ti 24981a750a1a bash
[root@24981a750a1a /]# ip route
default via 172.17.42.1 dev eth0
172.17.0.0/16 dev eth0 proto kernel scope link src 172.17.0.10
[root@24981a750a1a /]# ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
        valid_lft forever preferred_lft forever
22: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP
    link/ether 02:42:ac:11:00:0a brd ff:ff:ff:ff:ff:ff
    inet 172.17.0.10/16 scope global eth0
        valid_lft forever preferred_lft forever
    inet6 fe80::42:acff:fe11:a/64 scope link
        valid_lft forever preferred_lft forever
```

我们可以看到，默认停止的回环设备 `lo` 已经被启动，外面宿主机连接进来的 `Veth` 设备也被命名成了 `eth0`，并且已经配置了地址 `172.17.0.10`。

路由信息表包含一条到 `docker0` 的子网路由和一条到 `docker0` 的默认路由。

### 3) 查看容器启动后的情况（容器有端口映射）

下面，我们用带端口映射的命令启动 `registry`：

```
docker run --name register -d -p 1180:5000 registry
```

在启动后查看 `Iptables` 的变化：

```
# iptables-save
# Generated by iptables-save v1.4.21 on Thu Sep 24 18:45:13 2015
*nat
:PREROUTING ACCEPT [2:236]
:INPUT ACCEPT [0:0]
:OUTPUT ACCEPT [0:0]
:POSTROUTING ACCEPT [0:0]
:DOCKER - [0:0]
-A PREROUTING -m addrtype --dst-type LOCAL -j DOCKER
-A OUTPUT ! -d 127.0.0.0/8 -m addrtype --dst-type LOCAL -j DOCKER
-A POSTROUTING -s 172.17.0.0/16 ! -o docker0 -j MASQUERADE
-A POSTROUTING -s 172.17.0.19/32 -d 172.17.0.19/32 -p tcp -m tcp --dport 5000
-j MASQUERADE
-A DOCKER ! -i docker0 -p tcp -m tcp --dport 1180 -j DNAT --to-destination
172.17.0.19:5000
COMMIT
# Completed on Thu Sep 24 18:45:13 2015
# Generated by iptables-save v1.4.21 on Thu Sep 24 18:45:13 2015
*filter
:INPUT ACCEPT [54:4464]
:FORWARD ACCEPT [0:0]
:OUTPUT ACCEPT [41:5576]
:DOCKER - [0:0]
-A FORWARD -o docker0 -j DOCKER
-A FORWARD -o docker0 -m conntrack --ctstate RELATED,ESTABLISHED -j ACCEPT
-A FORWARD -i docker0 ! -o docker0 -j ACCEPT
-A FORWARD -i docker0 -o docker0 -j ACCEPT
-A DOCKER -d 172.17.0.19/32 ! -i docker0 -o docker0 -p tcp -m tcp --dport 5000
-j ACCEPT
COMMIT
# Completed on Thu Sep 24 18:45:13 2015
```

从新增的规则可以看出，`Docker` 服务在 `NAT` 和 `FILTER` 两个表内添加的两个 `DOCKER` 子链都是给端口映射用的。例如本例中我们需要把外面宿主机的 `1180` 端口映射到容器的 `5000` 端口。通过前面的分析我们知道，无论是宿主机接收到的还是宿主机本地协议栈发出的，目标地



址是本地 IP 地址的包都会经过 NAT 表中的 DOCKER 子链。Docker 为每一个端口映射都在这个链上增加了到实际容器目标地址和目标端口的转换。

经过这个 DNAT 的规则修改后的 IP 包，会重新经过路由模块的判断进行转发。由于目标地址和端口已经是容器的地址和端口，所以数据自然就送到了 docker0 上，从而送到对应的容器内部。

当然在 Forward 时，也需要在 Docker 子链中添加一条规则，如果目标端口和地址是指定容器的数据，则允许通过。

在 Docker 按照端口映射的方式启动容器时，主要的不同就是上述 Iptables 部分。而容器内部的路由和网络设备，都和不做端口映射时一样，没有任何变化。

#### 4) Docker 的网络局限

我们从 Docker 对 Linux 网络协议栈的操作可以看到，Docker 一开始没有考虑到多主机互联的网络解决方案。

Docker 一直以来的理念都是“简单为美”，几乎所有尝试 Docker 的人，都被它“用法简单，功能强大”的特性所吸引，这也是 Docker 迅速走红的一个原因。

我们都知道，虚拟化技术中最为复杂的部分就是虚拟化网络技术，即使是单纯的物理网络部分，也是一个门槛很高的技能领域，通常只被少数网络工程师所掌握，所以我们可以理解，结合了物理网络的虚拟网络技术会有多难了。在 Docker 之前，所有接触过 OpenStack 的人的心里都有一个难以释怀的阴影，那就是它的网络问题，于是，Docker 明智地避开这个“雷区”，让其他专业人员去用现有的虚拟化网络技术解决 Docker 主机的互联问题，以免让用户觉得 Docker 太难了，从而放弃学习和使用 Docker。

Docker 成名以后，重新开始重视网络解决方案，收购了一家 Docker 网络解决方案公司——Socketplane，原因在于这家公司的产品广受好评，但有趣的是 Socketplane 的方案就是以 Open vSwitch 为核心的，其还为 Open vSwitch 提供了 Docker 镜像，以方便部署程序。之后，Docker 开启了一个“宏伟”的虚拟化网络解决方案——Libnetwork，如图 3.24 所示是其概念图。

这个概念图没有了 IP，也没有了路由，已经颠覆了我们的网络常识了，对于不怎么懂网络的大多数人来说，它的确很有诱惑力，未来是否会对虚拟化网络的模型产生深远冲击我们还不得而知，但当前，它仅仅是 Docker 官方的一次“尝试”。

针对目前 Docker 的网络实现，Docker 使用的 Libnetwork 组件只是将 Docker 平台中的网络子系统模块化为一个独立库的简单尝试，离成熟和完善还有一段距离。

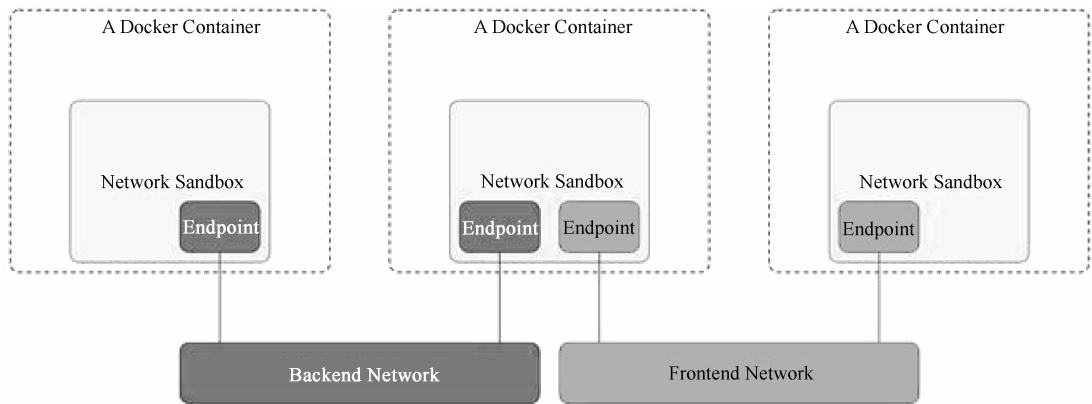


图 3.24 Libnetwork 概念图

### 3.7.4 Kubernetes 的网络实现

在实际的业务场景中，业务组件之间的关系十分复杂，特别是随着微服务理念逐步深入人心，应用部署的粒度更加细小和灵活。为了支持业务应用组件的通信联系，Kubernetes 网络的设计主要致力于解决以下场景。

- (1) 容器到容器之间的直接通信。
- (2) 抽象的 Pod 到 Pod 之间的通信。
- (3) Pod 到 Service 之间的通信。
- (4) 集群外部与内部组件之间的通信。

其中第 3 条、第 4 条我们在之前的章节里都讲述过，本节中我们对更为基础的第 1 条与第 2 条进行深入分析和讲解。

#### 1. 容器到容器的通信

在同一个 Pod 内的容器（Pod 内的容器是不会跨宿主机的）共享同一个网络命名空间，共享同一个 Linux 协议栈。所以对于网络的各类操作，就和它们在同一台机器上一样，它们甚至可以用 localhost 地址访问彼此的端口。

这么做的结果是简单、安全和高效，也能减少将已经存在的程序从物理机或者虚拟机移植到容器下运行的难度。在容器技术出来之前，其实大家早就积累了如何在一台机器上运行一组应用程序的经验，例如，如何让端口不冲突，以及如何让客户端发现它们等。

我们来看一下 Kubernetes 是如何利用 Docker 的网络模型的。

图 3.25 中的阴影部分就是在 Node 上运行着的一个 Pod 实例。在我们的例子中，容器就是图 3.25 中的容器 1 和容器 2。容器 1 和容器 2 共享了一个网络的命名空间，共享一个命名空间的结果就是它们好像在一台机器上运行似的，它们打开的端口不会有冲突，可以直接使用 Linux 的本地 IPC 进行通信（例如消息队列或者管道）。其实这 and 传统的一组普通程序运行的环境是完全一样的，传统的程序不需要针对网络做特别的修改就可以移植了。它们之间的互相访问只需要使用 localhost 就可以。例如，如果容器 2 运行的是 MySQL，那么容器 1 使用 localhost:3306 就能直接访问这个运行在容器 2 上的 MySQL 了。

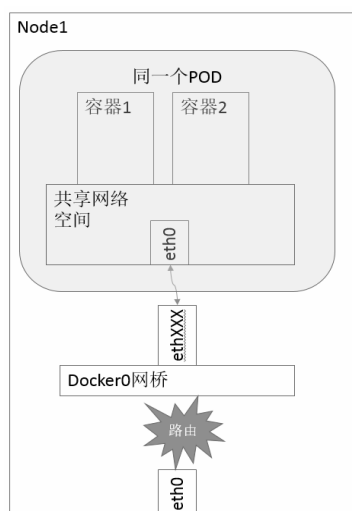


图 3.25 Kubernetes 的 Pod 网络模型

## 2. Pod 之间的通信

我们看了同一个 Pod 内的容器之间的通信情况，再看看 Pod 之间的通信情况。

每一个 Pod 都有一个真实的全局 IP 地址，同一个 Node 内的不同 Pod 之间可以直接采用对方 Pod 的 IP 地址通信，而且不需要使用其他发现机制，例如 DNS、Consul 或者 etcd。

Pod 容器既有可能在同一个 Node 上运行，也有可能在不同的 Node 上运行，所以通信也分为两类：同一个 Node 内的 Pod 之间的通信和不同 Node 上的 Pod 之间的通信。

### 1) 同一个 Node 内的 Pod 之间的通信

我们看一下同一个 Node 上的两个 Pod 之间的关系，如图 3.26 所示。

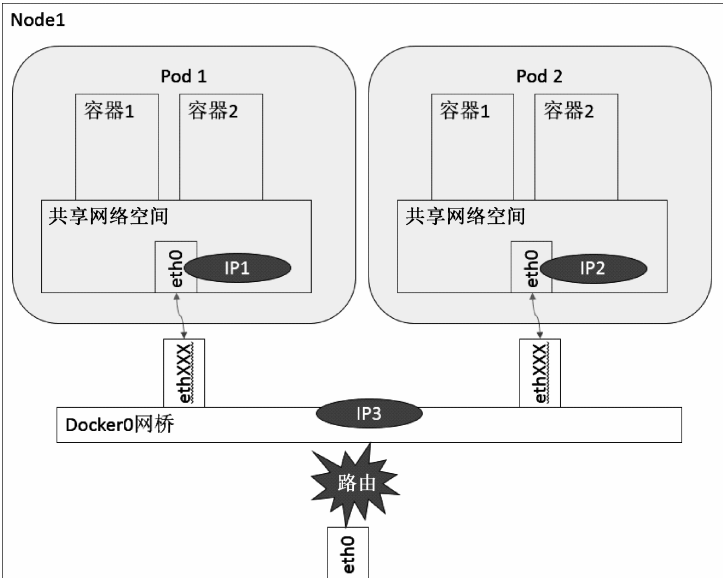


图 3.26 同一个 Node 内的 Pod 关系

可以看出,Pod1 和 Pod2 都是通过 Veth 连接在同一个 docker0 网桥上的,它们的 IP 地址 IP1、IP2 都是从 docker0 的网段上动态获取的,它们和网桥本身的 IP3 是同一个网段的。

另外,在 Pod1、Pod2 的 Linux 协议栈上,默认路由都是 docker0 的地址,也就是说所有非本地地址的网络数据,都会被默认发送到 docker0 网桥上,由 docker0 网桥直接中转。

综上所述,由于它们都关联在同一个 docker0 网桥上,地址段相同,所以它们之间是能直接通信的。

2) 不同 Node 上的 Pod 之间的通信

Pod 的地址是与 docker0 在同一个网段内的,我们知道 docker0 网段与宿主机网卡是两个完全不同的 IP 网段,并且不同 Node 之间的通信只能通过宿主机的物理网卡进行,因此要想实现位于不同 Node 上的 Pod 容器之间的通信,就必须想办法通过主机的这个 IP 地址来进行寻址和通信。

另一方面,这些动态分配且藏在 docker0 之后的所谓“私有”IP 地址也是可以找到的。Kubernetes 会记录所有正在运行 Pod 的 IP 分配信息,并将这些信息保存在 etcd 中(作为 Service 的 Endpoint)。这些私有 IP 信息对于 Pod 到 Pod 的通信也是十分重要的,因为我们的网络模型要求 Pod 到 Pod 使用私有 IP 进行通信。所以首先要知道这些 IP 是什么。

之前提到,Kubernetes 的网络对 Pod 的地址是平面的和直达的,所以这些 Pod 的 IP 规划也很重要,不能有冲突。只要没有冲突,我们就可以想办法在整个 Kubernetes 的集群中找到它。

综上所述，要想支持不同 Node 上的 Pod 之间的通信，就要达到两个条件：

(1) 在整个 Kubernetes 集群中对 Pod 的 IP 分配进行规划，不能有冲突；

(2) 找到一种办法，将 Pod 的 IP 和所在 Node 的 IP 关联起来，通过这个关联让 Pod 可以互相访问。

根据条件 1 的要求，我们需要在部署 Kubernetes 时，对 docker0 的 IP 地址进行规划，保证每一个 Node 上的 docker0 地址没有冲突。我们可以在规划后手工配置到每个 Node 上，或者做一个分配规则，由安装的程序自己去分配占用。例如 Kubernetes 的网络增强开源软件 Flannel 就能够管理资源池的分配。

根据条件 2 的要求，Pod 中的数据在发出时，需要有一个机制能够知道对方 Pod 的 IP 地址挂在哪个具体的 Node 上。也就是说先要找到 Node 对应宿主机的 IP 地址，将数据发送到这个宿主机的网卡上，然后在宿主机上将相应的数据转到具体的 docker0 上。一旦数据到达宿主机 Node，则那个 Node 内部的 docker0 便知道如何将数据发送到 Pod。如图 3.27 所示。

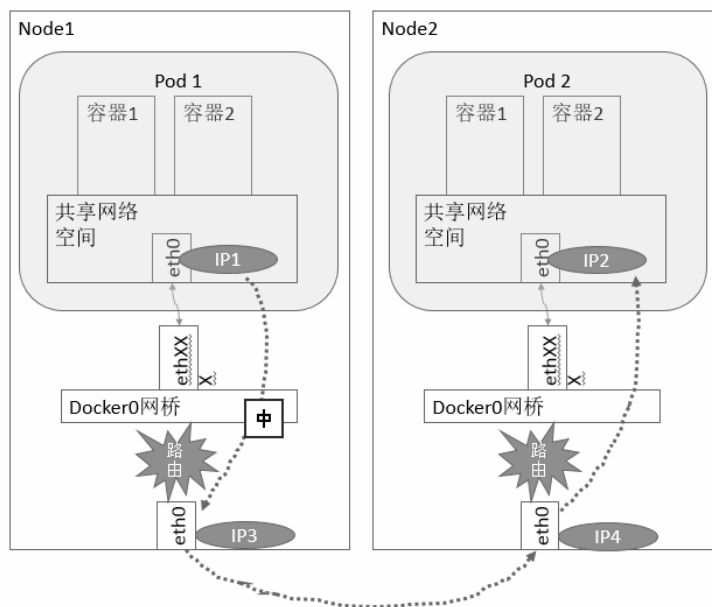


图 3.27 跨 Node 的 Pod 通信

在图 3.27 中，IP1 对应的是 Pod1，IP2 对应的是 Pod2。Pod1 在访问 Pod2 时，首先要将数据从源 Node 的 eth0 发送出去，找到并到达 Node2 的 eth0。也就是说先要从 IP3 到 IP4，之后才是 IP4 到 IP2 的递送。

在谷歌的 GCE 环境下，Pod 的 IP 管理（类似 docker0）、分配及它们之间的路由打通都是

由 GCE 完成的。Kubernetes 作为主要在 GCE 上面运行的框架，它的设计是假设底层已经具备这些条件，所以它分配完地址并将地址记录下来就完成了它的工作。在实际的 GCE 环境中，GCE 的网络组件会读取这些信息，实现具体的网络打通。

而在实际的生产中，因为安全、费用、合规等种种原因，Kubernetes 的客户不可能全部使用谷歌的 GCE 环境，所以在实际的私有云环境中，除了部署 Kubernetes 和 Docker，还需要额外的网络配置，甚至通过一些软件来实现 Kubernetes 对网络的要求。做到这些后，Pod 和 Pod 之间才能无差别地透明通信。

为了达到这个目的，开源界有不少应用来增强 Kubernetes、Docker 的网络，在后面的章节里会介绍几个常用的组件和它们的组网原理。

### 3.7.5 Pod 和 Service 网络实战

Docker 给我们带来了不同的网络模式，而 Kubernetes 也以一种不同的方式来解决这些网络模式的挑战，但其方式有些不太好理解，特别是对于刚开始接触 Kubernetes 的网络的开发者。我们在前面学习了 Kubernetes、Docker 的理论，本节将通过一个完整的实验，从部署一个 Pod 开始，一步一步地部署那些 Kubernetes 的组件，来剖析 Kubernetes 在网络层是如何实现及如何工作的。

这里使用虚拟机来完成实验。如果你要部署在物理机器上，或者部署在云服务商的环境下，则涉及的网络模型很可能稍微有所不同。不过，从网络角度来看，Kubernetes 的机制是类似且一致的。

好了，来看看我们的实验环境，如图 3.28 所示。

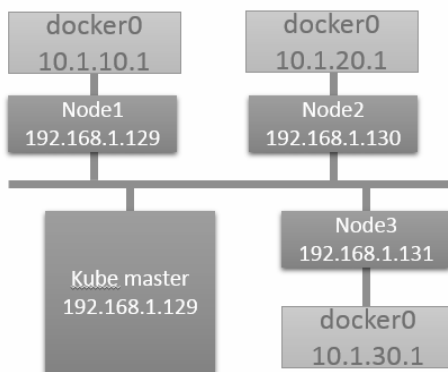


图 3.28 实验环境

Kubernetes 的网络模型要求每个 Node 上的容器都可以相互访问。

默认的 Docker 的网络模型提供了一个 IP 地址段是 172.17.0.0/16 的 docker0 网桥。每个容器都会在这个子网内获得 IP 地址，并且将 docker0 网桥的 IP 地址（172.17.42.1）作为其默认网关。需要注意的是 Docker 宿主机外面的网络不需要知道任何关于这个 172.17.0.0/16 的信息或者知道如何连接到它内部，因为 Docker 的宿主机针对容器发出的数据，在物理网卡地址后面都做了 IP 伪装 MASQUERADE（隐含 NAT）。也就是说，在网络上看到的任何容器数据流都来源于那台 Docker 节点的物理 IP 地址。这里所说的网络都是指连接这些主机的物理网络。

这个模型便于使用，但是并不完美，需要依赖端口映射的机制。

在 Kubernetes 的网络模型中，每台主机上的 docker0 网桥都是可以路由到的。也就是说，在部署了一个 Pod 时，在同一个集群内，那台主机的外面可以直接访问到那个 Pod，并不需要在那台物理主机上做端口映射。综上所述，你可以在网络层将 Kubernetes 的节点看作一个路由器。如果我们将实验环境改画成一个网络图，那么它看起来如图 3.29 所示。

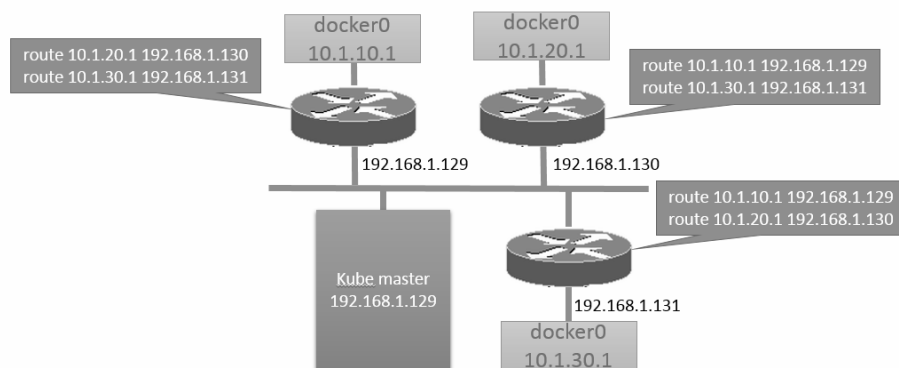


图 3.29 实验环境网络图

为了支持 Kubernetes 网络模型，我们采取了直接路由的方式来实现，在每个 Node 上配置相应的静态路由项，例如在 192.168.1.129 这个 Node 上配置了两个路由项：

```
#route add -net 10.1.20.0 netmask 255.255.255.0 gw 192.168.130
#route add -net 10.1.30.0 netmask 255.255.255.0 gw 192.168.131
```

这意味着，每一个新部署的容器都将使用这个 Node（docker0 的网桥 IP）作为它的默认网关。而这些 Node 节点（类似路由器）都有其他 docker0 的路由信息，这样它们就能够相互连通了。

接下来通过一些实际的案例，来看看 Kubernetes 在不同的场景下其网络部分到底做了什么事情。

## 第 1 步：部署一个 RC/Pod

部署的 RC/Pod 描述文件如下（frontend-controller.yaml）：

```
apiVersion: v1
kind: ReplicationController
metadata:
  name: frontend
  labels:
    name: frontend
spec:
  replicas: 1
  selector:
    name: frontend
  template:
    metadata:
      labels:
        name: frontend
    spec:
      containers:
        - name: php-redis
          image: kubeguide/guestbook-php-frontend
          env:
            - name: GET_HOSTS_FROM
              value: env
          ports:
            - containerPort: 80
              hostPort: 80
```

为了便于观察，我们假定在一个空的 Kubernetes 集群上运行，提前清理了所有 Replication Controller、Pod 和其他 Service：

```
# kubectl get rc
CONTROLLER  CONTAINER(S)  IMAGE(S)  SELECTOR  REPLICAS
#
# kubectl get services
NAME          LABELS          SELECTOR  IP(S)      PORT(S)
kubernetes    component=apiserver,provider=kubernetes <none>    20.1.0.1  443/TCP
#
# kubectl get pods
NAME          READY    STATUS    RESTARTS  AGE
```

让我们检查一下此时某个 Node 上的网络接口都有哪些。Node1 的状态是：

```
# ifconfig
docker0: flags=4099<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500
    inet 10.1.10.1 netmask 255.255.255.0 broadcast 10.1.10.255
    inet6 fe80::5484:7aff:fefe:9799 prefixlen 64 scopeid 0x20<link>
    ether 56:84:7a:fe:97:99 txqueuelen 0 (Ethernet)
```



```

RX packets 373245 bytes 170175373 (162.2 MiB)
RX errors 0 dropped 0 overruns 0 frame 0
TX packets 353569 bytes 353948005 (337.5 MiB)
TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

eno16777736: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500
    inet 192.168.1.129 netmask 255.255.255.0 broadcast 192.168.1.255
    inet6 fe80::20c:29ff:fe47:6e2c prefixlen 64 scopeid 0x20<link>
    ether 00:0c:29:47:6e:2c txqueuelen 1000 (Ethernet)
    RX packets 326552 bytes 286033393 (272.7 MiB)
    RX errors 0 dropped 0 overruns 0 frame 0
    TX packets 219520 bytes 31014871 (29.5 MiB)
    TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING> mtu 65536
    inet 127.0.0.1 netmask 255.0.0.0
    inet6 ::1 prefixlen 128 scopeid 0x10<host>
    loop txqueuelen 0 (Local Loopback)
    RX packets 24095 bytes 2133648 (2.0 MiB)
    RX errors 0 dropped 0 overruns 0 frame 0
    TX packets 24095 bytes 2133648 (2.0 MiB)
    TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

```

可以看出，有一个 **docker0** 网桥和一个本地地址的网络端口。现在部署一下我们在前面准备的 **RC/Pod** 配置文件，看看发生了什么：

```

# kubectl create -f frontend-controller.yaml
replicationcontrollers/frontend
#
# kubectl get pods
NAME                READY    STATUS    RESTARTS   AGE    NODE
frontend-4o1lg      1/1      Running   0           11s    192.168.1.130

```

可以看到一些有趣的事情。**Kubernetes** 为这个 Pod 找了一个主机 192.168.1.130 (Node2) 来运行它。另外，这个 Pod 还获得了一个在 Node2 的 **docker0** 网桥上的 IP 地址。我们登录到 Node2 上看看发生了什么事情：

```

# docker ps
CONTAINER ID        IMAGE               COMMAND             CREATED             STATUS             PORTS              NAMES
37b193a4c633       kubeguide/example-guestbook-php-redis   "/bin/sh -c /run.sh" 32 seconds ago     Up 26 seconds     k8s_php-redis.6ad3289e_frontend-n9n1m_development_813e2dd9-8149-11e5-823b-000c2921ba71_af6dd859
6dlb99cff4ae       google_containers/pause:latest           "/pause"             35 seconds ago     Up 28 seconds     0.0.0.0:80->80/tcp k8s_POD.855eeb3d_frontend-4t52y_development_813e3870-8149-11e5-823b-000c2921ba71_2b66f05e

```

在 Node2 上现在运行了两个容器，在我们的 **RC/Pod** 定义文件中仅仅包含了一个，那么这第 2 个是从哪里来的呢？第 2 个看起来运行的是一个叫作 **google\_containers/pause:latest** 的镜像，

而且这个容器已经有端口映射到它上面了，为什么是这样呢？让我们深入容器内部去看一下具体原因。使用 Docker 的“inspect”命令来查看容器的详细信息，特别要关注容器的网络模型。

```
# docker inspect 6d1b99cff4ae | grep NetworkMode
    "NetworkMode": "bridge",
# docker inspect 37b193a4c633 | grep NetworkMode
    "NetworkMode": "container:6d1b99cff4ae537689ce87d7528f4ba9dbb40ae711ecc0a5b3f7c39ff5e5e495",
```

有趣的结果是，在查看完每个容器的网络模型后，我们可以看到这样的配置：我们检查的第 1 个容器是运行了“google\_containers/pause:latest”镜像的容器，它使用了 Docker 默认的网络模型 bridge；而我们检查的第 2 个容器，也就是在 RC/Pod 中定义运行的 php-redis 容器，使用了非默认的网络配置和映射容器的模型，指定了映射目标容器为“google\_containers/ pause:latest”。

一起来仔细思考一下这个过程，为什么 Kubernetes 要这么做呢？首先，一个 Pod 内的所有容器都需要共用同一个 IP 地址，这就意味着一定要使用网络的容器映射模式。然而，为什么不能只启动第 1 个 Pod 中的容器，而将第 2 个 Pod 内的容器关联到第 1 个容器呢？我们认为 Kubernetes 从两个方面来考虑这个问题：首先，如果 Pod 有超过两个容器的话，则连接这些容器可能不容易；其次，后面的容器还要依赖第 1 个被关联的容器，如果第 2 个容器关联到第 1 个容器，且第 1 个容器死掉的话，第 2 个也将死掉。启动一个基础容器，然后将 Pod 内的所有容器都连接到它上面会更容易一些。因为我们只需要为基础的这个 Google\_containers/pause 容器执行端口映射规则，这也简化了端口映射的过程。所以我们启动 Pod 后的网络模型类似于图 3.30。

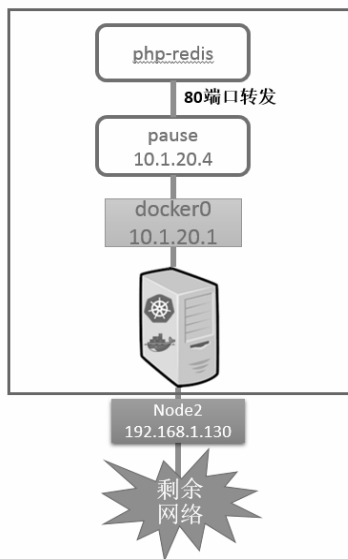


图 3.30 启动 Pod 后的网络模型

在这种情况下，实际 Pod 的 IP 数据流的网络目标都是这个 `google_containers/pause` 容器。图 3.30 有点儿取巧地显示了是 `google_containers/pause` 容器将端口 80 的流量转发给了相关的容器。而 `Pause` 只是逻辑上的，并没有真的这么做。实际上另外的 `Web` 容器直接监听了这些端口，和 `google_containers/pause` 容器共享了同一个网络堆栈。这就是为什么在 Pod 内部实际容器的端口映射都显示到 `google_containers/pause` 容器上了。我们可以通过 `docker port` 命令来检验一下：

```
# docker ps
CONTAINER ID          IMAGE
37b193a4c633         kubeguide/example-guestbook-php-redis
6d1b99cff4ae         google_containers/pause:latest
#
# docker port 6d1b99cff4ae
80/tcp -> 0.0.0.0:80
```

综上所述，`google_containers/pause` 容器实际上只是负责接管这个 Pod 的 Endpoint，并没有做更多的事情。那么 Node 呢？它需要将数据流传给 `google_containers/pause` 容器吗？我们来检查一下 `Iptables` 的规则，看看有什么发现：

```
# iptables-save
# Generated by iptables-save v1.4.21 on Thu Sep 24 17:15:01 2015
*nat
:PREROUTING ACCEPT [0:0]
:INPUT ACCEPT [0:0]
:OUTPUT ACCEPT [0:0]
:POSTROUTING ACCEPT [0:0]
:DOCKER - [0:0]
:KUBE-NODEPORT-CONTAINER - [0:0]
:KUBE-NODEPORT-HOST - [0:0]
:KUBE-PORTALS-CONTAINER - [0:0]
:KUBE-PORTALS-HOST - [0:0]
-A PREROUTING -m comment --comment "handle ClusterIPs; NOTE: this must be before
the NodePort rules" -j KUBE-PORTALS-CONTAINER
-A PREROUTING -m addrtype --dst-type LOCAL -j DOCKER
-A PREROUTING -m addrtype --dst-type LOCAL -m comment --comment "handle service
NodePorts; NOTE: this must be the last rule in the chain" -j KUBE-NODEPORT-CONTAINER
-A OUTPUT -m comment --comment "handle ClusterIPs; NOTE: this must be before the
NodePort rules" -j KUBE-PORTALS-HOST
-A OUTPUT ! -d 127.0.0.0/8 -m addrtype --dst-type LOCAL -j DOCKER
-A OUTPUT -m addrtype --dst-type LOCAL -m comment --comment "handle service
NodePorts; NOTE: this must be the last rule in the chain"
-A POSTROUTING -s 10.1.20.0/24 ! -o docker0 -j MASQUERADE
-A KUBE-PORTALS-CONTAINER -d 20.1.0.1/32 -p tcp -m comment --comment
"default/kubernetes:" -m tcp --dport 443 -j REDIRECT --to-ports 60339
-A KUBE-PORTALS-HOST -d 20.1.0.1/32 -p tcp -m comment --comment
"default/kubernetes:" -m tcp --dport 443 -j DNAT --to-destination 192.168.1.131:60339
```

```
COMMIT
# Completed on Thu Sep 24 17:15:01 2015
# Generated by iptables-save v1.4.21 on Thu Sep 24 17:15:01 2015
*filter
:INPUT ACCEPT [1131:377745]
:FORWARD ACCEPT [0:0]
:OUTPUT ACCEPT [1246:209888]
:DOCKER - [0:0]
-A FORWARD -o docker0 -j DOCKER
-A FORWARD -o docker0 -m conntrack --ctstate RELATED,ESTABLISHED -j ACCEPT
-A FORWARD -i docker0 ! -o docker0 -j ACCEPT
-A FORWARD -i docker0 -o docker0 -j ACCEPT
-A DOCKER -d 172.17.0.19/32 ! -i docker0 -o docker0 -p tcp -m tcp --dport 5000
-j ACCEPT
COMMIT
# Completed on Thu Sep 24 17:15:01 2015
```

上面的这些规则并没有应用到我们刚刚定义的 Pod 上。当然，Kubernetes 会给每一个 Kubernetes 的节点提供一些默认的服务，上面的规则就是 Kubernetes 的默认服务需要的。关键是，我们没有看到任何 IP 伪装的规则，并且没有任何指向 Pod 10.1.20.4 的内部方向的端口映射。

## 第 2 步：发布一个服务

我们已经了解了 Kubernetes 如何处理最基本的元素即 Pod 的连接问题，接下来看一下它是如何处理 Service 的。Service 允许我们在多个 Pod 之间抽象一些服务，而且服务可以通过提供在同一个 Service 的多个 Pod 之间的负载均衡机制来支持水平扩展。我们再次将环境初始化，删除刚刚创建的 RC/Pod 来确保集群是空的：

```
# kubectl stop rc frontend
replicationcontroller/frontend
#
# kubectl get rc
CONTROLLER  CONTAINER(S)  IMAGE(S)  SELECTOR  REPLICAS
#
# kubectl get services
NAME          LABELS                                SELECTOR  IP(S)      PORT(S)
kubernetes    component=apiserver,provider=kubernetes  <none>    20.1.0.1
443/TCP
#
# kubectl get pods
NAME          READY  STATUS  RESTARTS  AGE
```

然后准备一个名称为 frontend 的 Service 配置文件：

```
apiVersion: v1
kind: Service
```

```

metadata:
  name: frontend
  labels:
    name: frontend
spec:
  ports:
    - port: 80
#   nodePort: 30001
  selector:
    name: frontend
# type:
#   NodePort

```

然后在 Kubernetes 集群中定义这个服务：

```

# kubectl create -f frontend-service.yaml
services/frontend
# kubectl get services
NAME          LABELS              SELECTOR              IP(S)              PORT(S)
frontend      name=frontend       name=frontend         20.1.244.75        80/TCP
kubernetes    component=apiserver,provider=kubernetes <none>              20.1.0.1
443/TCP

```

服务正确创建后，可以看到 Kubernetes 集群已经为这个服务分配了一个虚拟 IP 地址 20.1.244.75，这个 IP 地址是在 Kubernetes 的 Portal Network 中分配的。而这个 Portal Network 的地址范围则是我们在 Kubmaster 上启动 API 服务进程时，使用 `--service-cluster-ip-range=xx` 命令行参数指定的：

```

# cat /etc/kubernetes/apiserver
.....
# Address range to use for services
KUBE_SERVICE_ADDRESSES="--service-cluster-ip-range=20.1.0.0/16"
.....

```

这个 IP 段可以是任何段，只要不和 `docker0` 或者物理网络的子网冲突就可以。选择任意其他网段的原因是这个网段将不会在物理网络和 `docker0` 网络上进行路由。这个 Portal Network 针对每一个 Node 都有局部的特殊性，实际上它存在的意义是让容器的流量都指向默认网关（也就是 `docker0` 网桥）。在继续实验前，先登录到 Node1 上看一下我们定义服务后发生了什么变化。首先检查一下 Iptables/Netfilter 的规则：

```

# iptables-save
.....
-A KUBE-PORTALS-CONTAINER -d 20.1.244.75/32 -p tcp -m comment --comment "default/frontend:" -m tcp --dport 80 -j REDIRECT --to-ports 59528
-A KUBE-PORTALS-HOST -d 20.1.244.75/32 -p tcp -m comment --comment "default/kubernetes:" -m tcp --dport 80 -j DNAT --to-destination 192.168.1.131:59528
.....

```

第 1 行是挂在 **PREROUTING** 链上的端口重定向规则，所有的进流量如果满足 20.1.244.75:80，则都会被重定向到端口 33761。第 2 行是挂在 **OUTPUT** 链上的目标地址 NAT，做了和上述第 1 行规则类似的工作，但针对的是当前主机生成的外出流量。所有主机生成的流量都需要使用这个 **DNAT** 规则来处理。简而言之，这两个规则使用了不同的方式做了类似的事情，就是将所有从节点生成的发送给 20.1.244.75:80 的流量重定向到本地的 33761 端口。

至此为止，目标为 **Service IP** 地址和端口的任何流量都将被重定向到本地的 33761 端口上。这个端口连到哪里去了呢？这就到了 **kube-proxy** 发挥作用的地方了。这个 **kube-proxy** 服务给每一个新创建的服务关联了一个随机的端口号，并且监听那个特定的端口，为服务创建相关的负载均衡对象。在我们的实验中，随机生成的端口刚好是 33761。通过监控 **Node1** 上的 **Kubernetes-Service** 的日志，在创建服务时，我们可以看到下面的记录：

```
2612 proxier.go:413] Opened iptables from-containers portal for service "default/
frontend:" on TCP 20.1.244.75:80
2612 proxier.go:424] Opened iptables from-host portal for service "default/
frontend:" on TCP 20.1.244.75:80
```

现在我们知道，所有流量都被导入 **kube-proxy** 中。现在我们需要它完成一些负载均衡的工作。创建 **Replication Controller** 并观察结果，下面是 **Replication Controller** 的配置文件：

```
apiVersion: v1
kind: ReplicationController
metadata:
  name: frontend
  labels:
    name: frontend
spec:
  replicas: 3
  selector:
    name: frontend
  template:
    metadata:
      labels:
        name: frontend
    spec:
      containers:
      - name: php-redis
        image: kubeguide/example-guestbook-php-redis
        env:
        - name: GET_HOSTS_FROM
          value: env
        ports:
        - containerPort: 80
#       hostPort: 80
```

在集群发布上述配置文件后，等待并观察，确保所有 Pod 都运行起来了：

```
# kubectl create -f frontend-controller.yaml
replicationcontrollers/frontend
#
# kubectl get pods -o wide
```

NAME	READY	STATUS	RESTARTS	AGE	NODE
frontend-64t8q	1/1	Running	0	5s	192.168.1.130
frontend-dzqve	1/1	Running	0	5s	192.168.1.131
frontend-x5dwy	1/1	Running	0	5s	192.168.1.129

现在所有的 Pod 都运行起来了，Service 将会对匹配到标签为“name=frontend”的所有 Pod 进行负载分发。因为 Service 的选择匹配所有的这些 Pod，所以我们的负载均衡将会对这 3 个 Pod 进行分发。现在的实验环境如图 3.31 所示。

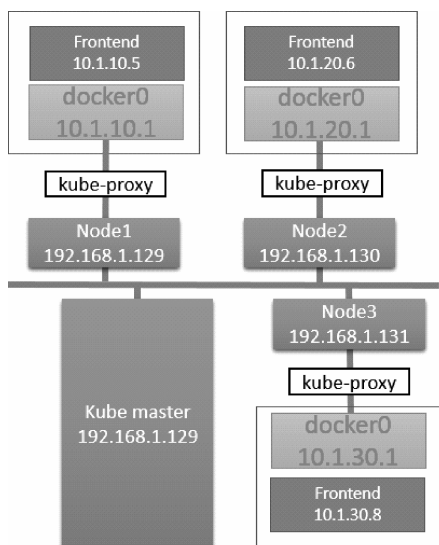


图 3.31 现在的实验环境

Kubernetes 的 kube-proxy 看起来只是一个夹层，但实际上它只是在 Node 上运行的一个服务。上述重定向规则的结果就是针对目标地址为服务 IP 的流量，将 Kubernetes 的 kube-proxy 变成了一个中间的夹层。

为了查看具体的重定向动作，我们会使用 tcpdump 来进行网络抓包操作。首先，安装 tcpdump：

```
yum -y install tcpdump
```

安装完成后，登录 Node1，运行 tcpdump 命令：

```
tcpdump -nn -q -i eno16777736 port 80
```

需要捕获物理服务器以太网接口的数据包，Node1 机器上的以太网接口名字叫作 eno16777736。

再打开第 1 个窗口运行第 2 个 `tcpdump` 程序，不过我们需要一些额外的信息去运行它，即挂在在 `docker0` 桥上的虚拟网卡 `Veth` 的名字。我们看到只有一个 `frontend` 容器在 `Node1` 主机上运行，所以可以使用简单的“`ip addr`”命令来查看唯一的“`Veth`”网络接口：

```
# ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
        valid_lft forever preferred_lft forever
2: eno16777736: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast
state UP qlen 1000
    link/ether 00:0c:29:47:6e:2c brd ff:ff:ff:ff:ff:ff
    inet 192.168.1.129/24 brd 192.168.1.255 scope global eno16777736
        valid_lft forever preferred_lft forever
    inet6 fe80::20c:29ff:fe47:6e2c/64 scope link
        valid_lft forever preferred_lft forever
3: docker0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc noqueue state DOWN
    link/ether 56:84:7a:fe:97:99 brd ff:ff:ff:ff:ff:ff
    inet 10.1.10.1/24 brd 10.1.10.255 scope global docker0
        valid_lft forever preferred_lft forever
    inet6 fe80::5484:7aff:fefe:9799/64 scope link
        valid_lft forever preferred_lft forever
12: veth0558bfa: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue master
docker0 state UP
    link/ether 86:82:e5:c8:5a:9a brd ff:ff:ff:ff:ff:ff
    inet6 fe80::8482:e5ff:fec8:5a9a/64 scope link
        valid_lft forever preferred_lft forever
```

复制这个接口的名字，在第 2 个窗口中运行 `tcpdump` 命令：

```
tcpdump -nn -q -i veth0558bfa host 20.1.244.75
```

同时运行这两个命令，并且将窗口并排放置，以便同时看到两个窗口的输出：

```
# tcpdump -nn -q -i eno16777736 port 80
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on eno16777736, link-type EN10MB (Ethernet), capture size 65535 bytes

# tcpdump -nn -q -i veth0558bfa host 20.1.244.75
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on veth0558bfa, link-type EN10MB (Ethernet), capture size 65535 bytes
```

好了，我们已经在同时捕获两个接口的网络包了。这时再启动第 3 个窗口，运行一个“`docker exec`”命令来连接到我们的“`frontend`”的容器内部（你可以先执行 `docker ps` 来获得这个容器的 ID）：

```
# docker ps
```



CONTAINER ID	IMAGE	.....
268ccdfb9524	kubeguide/example-guestbook-php-redis	.....
6a519772b27e	google_containers/pause:latest	.....

执行命令进入容器内部:

```
#docker exec -it 268ccdfb9524 bash
# docker exec -it 268ccdfb9524 bash
root@frontend-x5dwy:/#
```

一旦进入运行的容器内部,我们就可以通过 Pod 的 IP 地址来访问服务了。使用 curl 来尝试访问服务:

```
curl 20.1.244.75
```

在使用 curl 访问服务时,将在抓包的两个窗口内看到:

```
20:19:45.208948 IP 192.168.1.129.57452 > 10.1.30.8.8080: tcp 0
20:19:45.209005 IP 10.1.30.8.8080 > 192.168.1.129.57452: tcp 0
20:19:45.209013 IP 192.168.1.129.57452 > 10.1.30.8.8080: tcp 0
20:19:45.209066 IP 10.1.30.8.8080 > 192.168.1.129.57452: tcp 0
```

```
20:19:45.209227 IP 10.1.10.5.35225 > 20.1.244.75.80: tcp 0
20:19:45.209234 IP 20.1.244.75.80 > 10.1.10.5.35225: tcp 0
20:19:45.209280 IP 10.1.10.5.35225 > 20.1.244.75.80: tcp 0
20:19:45.209336 IP 20.1.244.75.80 > 10.1.10.5.35225: tcp 0
```

这些信息说明了什么问题呢? 让我们在网络图上用实线标出第 1 个窗口中网络抓包信息的含义(物理网卡上的网络流量),并用虚线标出第 2 个窗口中网络抓包信息的含义(docker0 网桥上的网络流量),如图 3.32 所示。

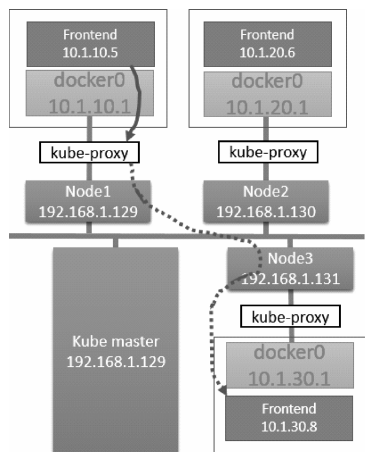


图 3.32 数据流动情况图 1

注意,图 3.32 中,虚线绕过了 Node3 的 kube-proxy,这么做是因为 Node3 上的 kube-proxy

没有参与这次网络交互。换句话说，Node1 的 kube-proxy 服务和负载均衡到的 Pod 进行网络交互。

在查看第 2 个捕获包的窗口时，我们能够站在容器的视角看这些流量。首先，容器尝试使用 20.1.244.75:80 打开 TCP 的 Socket 连接。同时，我们可以看到从服务地址 20.1.244.75 返回的数据。从容器的视角来看，整个交互过程都是在服务之间进行的。但是在查看一个捕获包的窗口时（上面的窗口），我们可以看到物理机之间的数据交互，可以看到一个 TCP 连接从 Node1 的物理地址（192.168.1.129）发出，直接连接到运行 Pod 的主机 Node3（192.168.1.131）。总而言之，Kubernetes 的 kube-proxy 作为一个全功能的代理服务器管理了两个独立的 TCP 连接：一个是从容器到 kube-proxy；另一个是从 kube-proxy 到负载均衡的目标 Pod。

如果我们清理一下捕获的记录，再次运行 curl，则还可以看到网络流量被负载均衡转发到另一个节点 Node2 上了。

```
20:19:45.208948 IP 192.168.1.129.57485 > 10.1.20.6.8080: tcp 0
20:19:45.209005 IP 10.1.20.6.8080 > 192.168.1.129.57485: tcp 0
20:19:45.209013 IP 192.168.1.129.57485 > 10.1.20.6.8080: tcp 0
20:19:45.209066 IP 10.1.20.6.8080 > 192.168.1.129.57485: tcp 0

20:19:45.209227 IP 10.1.10.5.38026 > 20.1.244.75.80: tcp 0
20:19:45.209234 IP 20.1.244.75.80 > 10.1.10.5.38026: tcp 0
20:19:45.209280 IP 10.1.10.5.38026 > 20.1.244.75.80: tcp 0
20:19:45.209336 IP 20.1.244.75.80 > 10.1.10.5.38026: tcp 0
```

这一次，Kubernetes 的 Proxy 将选择运行在 Node2（10.1.20.1）上面的 Pod 作为负载均衡的目的。网络流动图如图 3.33 所示。

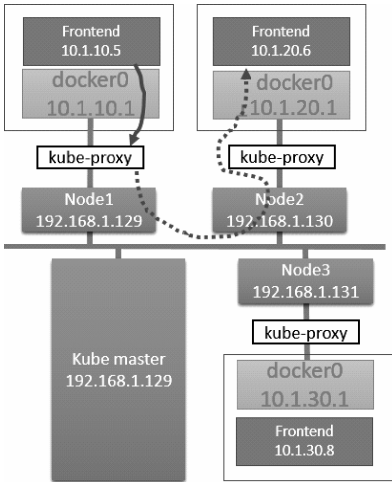


图 3.33 数据流动情况图 2

到这里，你肯定已经知道另外一个可能的负载均衡的路由结果了吧。

### 3.7.6 CNI 网络模型

随着容器技术在企业生产系统的逐步落地，用户对于容器云的网络特性要求也越来越高。跨主机容器间网络互通已经成为最基本的要求，更高的要求包括容器固定 IP 地址、一个容器多个 IP 地址、多个子网隔离、ACL 控制策略、与 SDN 集成等。目前主流的容器网络模型主要有 Docker 公司提出的 Container Network Model (CNM) 模型和 CoreOS 公司提出的 Container Network Interface (CNI) 模型。

#### 1. CNM 模型

CNM 模型是由 Docker 公司提出的容器网络模型，现在已经被 Cisco Contiv、Kuryr、Open Virtual Networking (OVN)、Project Calico、VMware、Weave 和 Plumgrid 等项目所采纳。另外，Weave、Project Calico、Kuryr 和 Plumgrid 等项目也为 CNM 提供网络插件的具体实现。

CNM 模型主要通过 Network Sandbox、Endpoint 和 Network 这 3 个组件进行实现，如图 3.34 所示。

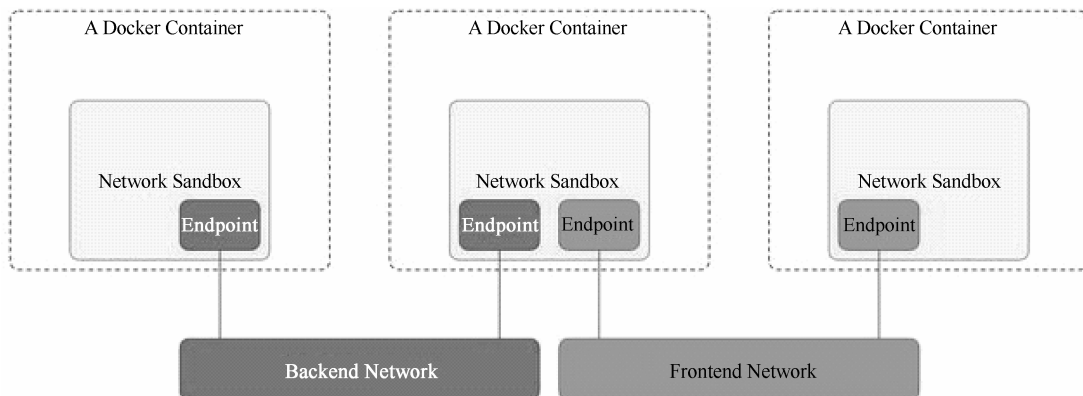


图 3.34 CNM 模型

- ◎ **Network Sandbox:** 容器内部的网络栈，包括网络接口、路由表、DNS 等配置的管理。Sandbox 可用 Linux 网络命名空间、FreeBSD Jail 等机制进行实现。一个 Sandbox 可以包含多个 Endpoint。
- ◎ **Endpoint:** 用于将容器内的 Sandbox 与外部网络相连的网络接口。可以使用 veth 对、Open vSwitch 的内部 port 等技术进行实现。一个 Endpoint 仅能够加入一个 Network。

- ◎ **Network**：可以直接互连的 Endpoint 的集合。可以通过 Linux 网桥、VLAN 等技术进行实现。一个 Network 包含多个 Endpoint。

## 2. CNI 模型

CNI 是由 CoreOS 公司提出的另一种容器网络规范，现在已经被 Kubernetes、rkt、Apache Mesos、Cloud Foundry 和 Kurma 等项目采纳。另外，Contiv Networking、Project Calico、Weave、SR-IOV、Cilium、Infoblox、Multus、Romana、Plumgrid 和 Midokura 等项目也为 CNI 提供网络插件的具体实现。图 3.35 描述了容器运行环境与各种网络插件通过 CNI 进行连接的模型。

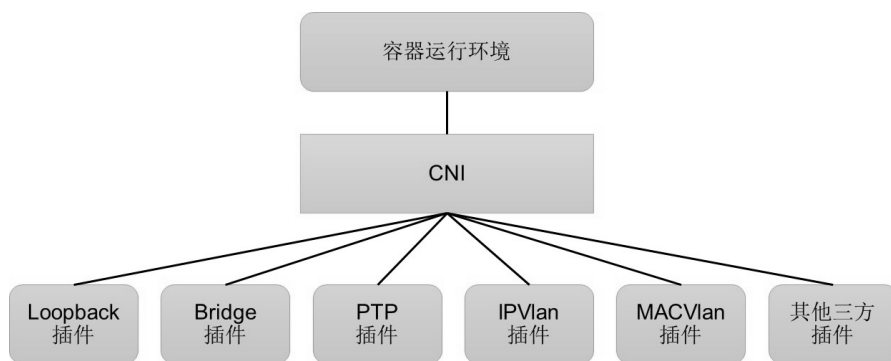


图 3.35 CNI 模型

CNI 定义的是容器运行环境与网络插件之间的简单接口规范，通过一个 JSON Schema 定义 CNI 插件提供的输入和输出参数。一个容器可以通过绑定多个网络插件加入多个网络中。

本节将对 Kubernetes 如何实现 CNI 模型进行详细说明。

### 1) CNI 规范概述

CNI 提供了一种应用容器的插件化网络解决方案，定义对容器网络进行操作和配置的规范，通过插件（plugin）的形式对 CNI 接口进行实现。CNI 是由 rkt Networking Proposal 发展而来的，试图提供一种普适的容器网络解决方案。CNI 仅关注在创建容器时分配网络资源，和在销毁容器时删除网络资源，这使得 CNI 规范非常轻巧，易于实现，得到了广泛的支持。

CNI 模型中只涉及两个概念：容器和网络。

- ◎ **容器（Container）**：容器是拥有独立 Linux 网络命名空间的环境，例如使用 Docker 或 rkt 创建的容器。关键之处是容器需要拥有自己的 Linux 网络命名空间，这是加入网络的必要条件。
- ◎ **网络（Network）**：网络表示可以互连的一组实体，这些实体拥有各自独立、唯一的 IP 地址。这些实体可以是容器、物理机或者其他网络设备（比如路由器）等。

对容器网络的设置和操作都通过插件（Plugin）进行具体实现，CNI 插件包括两种类型：CNI Plugin 和 IPAM(IP Address Management)Plugin。CNI Plugin 负责为容器配置网络资源，IPAM Plugin 负责对容器的 IP 地址进行分配和管理。IPAM Plugin 作为 CNI Plugin 的一部分，与 CNI Plugin 一起工作。

## 2) CNI Plugin 插件详解

CNI Plugin 包括 3 个基本接口的定义：添加（Add Container to Network）、删除（Delete Container from Network）和版本查询（Report Version）。这些接口的具体实现要求插件提供一个可执行的程序，在容器网络添加或删除时进行调用，以完成具体的操作。

（1）添加：将容器添加到某个网络。主要过程为在 Container Runtime 创建容器时，先创建好容器内的网络命名空间（Network Namespace），然后调用 CNI 插件为该 netns 进行网络配置，最后启动容器内的进程。

添加接口的参数如下。

- ◎ Version: CNI 版本号。
- ◎ Container ID: 容器 ID。
- ◎ Network namespace path: 容器的网络命名空间路径，例如/proc/[pid]/ns/net。
- ◎ Network configuration: 网络配置 JSON 文档，用于描述容器待加入的网络。
- ◎ Extra arguments: 其他参数，提供基于容器的 CNI 插件简单配置机制。
- ◎ Name of the interface inside the container: 容器内的网卡名。

返回的信息如下。

- ◎ Interfaces list: 网卡列表，根据 plugin 的实现，可能包括 Sandbox Interface 名称、主机 Interface 名称、每个 Interface 的地址等信息。
- ◎ IPs assigned to the interface: IPv4 或者 IPv6 地址、网关地址、路由信息等。
- ◎ DNS information: DNS 相关的信息。

（2）删除：容器销毁时将容器从某个网络中删除。

删除接口的参数如下。

- ◎ Version: CNI 版本号。
- ◎ Container ID: 容器 ID。
- ◎ Network namespace path: 容器的网络命名空间路径，例如/proc/[pid]/ns/net。
- ◎ Network configuration: 网络配置 JSON 文档，用于描述容器待加入的网络。

- ◎ **Extra arguments:** 其他参数，提供基于容器的 CNI 插件简单配置机制。
- ◎ **Name of the interface inside the container:** 容器内的网卡名。

(3) 版本查询：查询网络插件支持的 CNI 规范版本号。

无参数，返回值为网络插件支持的 CNI 规范版本号，例如：

```
{
  "cniVersion": "0.3.1", // the version of the CNI spec in use for this output
  "supportedVersions": [ "0.1.0", "0.2.0", "0.3.0", "0.3.1" ] // the list of CNI
spec versions that this plugin supports
}
```

CNI 插件应能够支持通过环境变量和标准输入传入参数。可执行文件通过“网络配置参数”中的 **type** 字段标识的文件名在环境变量“**CNI\_PATH**”设定的路径下进行查找。一旦找到，容器运行时将调用该可执行程序，并传入以下环境变量和网络配置参数，供该插件完成容器网络资源和参数的设置。

环境变量参数如下。

- ◎ **CNI\_COMMAND:** 接口方法，包括 ADD、DEL 和 VERSION。
- ◎ **CNI\_CONTAINERID:** 容器 ID。
- ◎ **CNI\_NETNS:** 容器的网络命名空间路径，例如/proc/[pid]/ns/net。
- ◎ **CNI\_IFNAME:** 待设置的网络接口名称。
- ◎ **CNI\_ARGS:** 其他参数，为 **key=value** 格式，多个参数之间用分号分隔，例如“**FOO=BAR;ABC=123**”。
- ◎ **CNI\_PATH:** 可执行文件查找路径，可以设置多个。

网络配置参数则由一个 JSON 报文组成，以标准输入（stdin）的方式传递给可执行程序。

网络配置参数如下。

- ◎ **cniVersion (string):** CNI 版本号。
- ◎ **name (string):** 网络名称，应在一个管理域内唯一。
- ◎ **type (string):** CNI 插件可执行文件的名称。
- ◎ **args (dictionary):** 其他参数。
- ◎ **ipMasq (boolean):** 是否设置 IP Masquerade（需插件支持），适用于主机可作为网关的环境中。
- ◎ **ipam:** IP 地址管理的相关配置。

- `type (string)`: IPAM 可执行的文件名。
- ◎ `dns`: DNS 服务的相关配置。
  - `nameservers (list of strings)`: 名字服务器列表, 可以使用 IPv4 或 IPv6 地址。
  - `domain (string)`: 本地域名, 用于短主机名查询。
  - `search (list of strings)`: 按优先级排序的域名查询列表。
  - `options (list of strings)`: 传递给 `resolver` 的选项列表。

下例定义了一个名为“dbnet”的网络配置参数, IPAM 使用“host-local”进行设置。

```
{
  "cniVersion": "0.3.1",
  "name": "dbnet",
  "type": "bridge",
  "bridge": "cni0",
  "ipam": {
    "type": "host-local",
    "subnet": "10.1.0.0/16",
    "gateway": "10.1.0.1"
  },
  "dns": {
    "nameservers": [ "10.1.0.1" ]
  }
}
```

### 3) IPAM Plugin 插件详解

为了减轻 CNI Plugin 对 IP 地址管理的负担, CNI 规范中设置了一个新的插件专门用于管理容器的 IP 地址 (还包括网关、路由等信息), 被称为 IPAM Plugin。通常由 CNI Plugin 在运行时自动调用 IPAM Plugin 完成容器 IP 地址的分配。

IPAM Plugin 负责为容器分配 IP 地址、网关、路由和 DNS, 典型的实现包括 `host-local` 和 `dhcp`。与 CNI Plugin 类似, IPAM 插件也通过可执行程序完成 IP 地址分配的具体操作。IPAM 可执行程序也处理传递给 CNI 插件的环境变量和通过标准输入 (`stdin`) 传入的网络配置参数。

如果成功完成了容器 IP 地址的分配, 则 IPAM 插件应该通过标准输出 (`stdout`) 返回以下 JSON 报文:

```
{
  "cniVersion": "0.3.1",
  "ips": [
    {
      "version": "<4-or-6>",
      "address": "<ip-and-prefix-in-CIDR>",
      "gateway": "<ip-address-of-the-gateway>" (optional)
    }
  ]
}
```

```
    },
    ...
  ],
  "routes": [                                     (optional)
    {
      "dst": "<ip-and-prefix-in-cidr>",
      "gw": "<ip-of-next-hop>"                     (optional)
    },
    ...
  ]
  "dns": {
    "nameservers": <list-of-nameservers>           (optional)
    "domain": <name-of-local-domain>               (optional)
    "search": <list-of-search-domains>            (optional)
    "options": <list-of-options>                   (optional)
  }
}
```

其中包括 `ips`、`routes` 和 `dns` 三段内容。

- ◎ `ips` 段：分配给容器的 IP 地址（也可能包括网关）。
- ◎ `routes` 段：路由规则记录。
- ◎ `dns` 段：DNS 相关的信息。

#### 4) 多网络插件

很多情况下，一个容器需要连接多个网络，CNI 规范支持为一个容器运行多个 CNI Plugin 来实现这个目标。多个网络插件将按照网络配置列表中的顺序执行，并将前一个网络配置的执行结果传递给后面的网络配置。多网络配置用 JSON 报文进行配置，包括如下信息。

- ◎ `cniVersion` (string)：CNI 版本号。
- ◎ `name` (string)：网络名称，应在一个管理域内唯一，将用于下面的所有 plugin。
- ◎ `plugins` (list)：网络配置列表。

下例定义了两个网络配置参数，分别作用于两个插件，第 1 个为 `bridge`，第 2 个为 `tuning`。CNI 将首先执行第 1 个 `bridge` 插件设置容器的网络，然后执行第 2 个 `tuning` 插件：

```
{
  "cniVersion": "0.3.1",
  "name": "dbnet",
  "plugins": [
    {
      "type": "bridge",
      // type (plugin) specific
      "bridge": "cni0",

```



```

// args may be ignored by plugins
"args": {
  "labels" : {
    "appVersion" : "1.0"
  }
},
"ipam": {
  "type": "host-local",
  // ipam specific
  "subnet": "10.1.0.0/16",
  "gateway": "10.1.0.1"
},
"dns": {
  "nameservers": [ "10.1.0.1" ]
}
},
{
  "type": "tuning",
  "sysctl": {
    "net.core.somaxconn": "500"
  }
}
]
}

```

容器运行时，在执行第 1 个 bridge 插件时，网络配置参数将被设置为：

```

{
  "cniVersion": "0.3.1",
  "name": "dbnet",
  "type": "bridge",
  "bridge": "cni0",
  "args": {
    "labels" : {
      "appVersion" : "1.0"
    }
  },
  "ipam": {
    "type": "host-local",
    // ipam specific
    "subnet": "10.1.0.0/16",
    "gateway": "10.1.0.1"
  },
  "dns": {
    "nameservers": [ "10.1.0.1" ]
  }
}

```

之后，在执行第 2 个 **tuning** 插件时，网络配置参数将被设置为：

```
{
  "cniVersion": "0.3.1",
  "name": "dbnet",
  "type": "tuning",
  "sysctl": {
    "net.core.somaxconn": "500"
  },
  "prevResult": {
    "ip4": {
      "ip": "10.1.0.3/16",
      "gateway": "10.1.0.1",
    },
    "dns": {
      "nameservers": [ "10.1.0.1" ]
    }
  }
}
```

其中 **prevResult** 字段内包含的信息为上一个 **bridge** 插件执行的结果。

在删除多个 CNI Plugin 时，则以逆序执行删除操作，以上例为例，将先删除 **tuning** 插件的网络配置：

```
{
  "cniVersion": "0.3.1",
  "name": "dbnet",
  "type": "tuning",
  "sysctl": {
    "net.core.somaxconn": "500"
  }
}
```

然后删除 **bridge** 插件的网络配置：

```
{
  "cniVersion": "0.3.1",
  "name": "dbnet",
  "type": "bridge",
  "bridge": "cni0",
  "args": {
    "labels" : {
      "appVersion" : "1.0"
    }
  },
  "ipam": {
    "type": "host-local",
    // ipam specific
  }
}
```

```

    "subnet": "10.1.0.0/16",
    "gateway": "10.1.0.1"
  },
  "dns": {
    "nameservers": [ "10.1.0.1" ]
  }
}

```

### 5) 命令返回信息说明

对 ADD 或 DELETE 操作，返回码为 0 表示执行成功，非 0 表示失败，并以 JSON 报文的格式通过标准输出 (stdout) 返回操作的结果。

以 ADD 操作为例，成功将容器添加到网络的结果将返回以下 JSON 报文。其中 ips、routes 和 dns 段的信息应该与 IPAM Plugin (IPAM Plugin 的说明详见下节) 返回的结果相同，重要的是 interfaces 段，应通过 CNI Plugin 进行设置并返回。

```

{
  "cniVersion": "0.3.1",
  "interfaces": [                                (this key omitted by IPAM plugins)
    {
      "name": "<name>",
      "mac": "<MAC address>",    (required if L2 addresses are meaningful)
      "sandbox": "<netns path or hypervisor identifier>" (required for
container/hypervisor interfaces, empty/omitted for host interfaces)
    }
  ],
  "ips": [
    {
      "version": "<4-or-6>",
      "address": "<ip-and-prefix-in-CIDR>",
      "gateway": "<ip-address-of-the-gateway>", (optional)
      "interface": <numeric index into 'interfaces' list>
    },
    ...
  ],
  "routes": [                                    (optional)
    {
      "dst": "<ip-and-prefix-in-cidr>",
      "gw": "<ip-of-next-hop>"                    (optional)
    },
    ...
  ],
  "dns": {
    "nameservers": <list-of-nameservers>          (optional)
    "domain": <name-of-local-domain>              (optional)
  }
}

```

```
    "search": <list-of-additional-search-domains>      (optional)
    "options": <list-of-options>                      (optional)
  }
}
```

接口调用失败时，返回码不为 0，应通过标准输出（stdout）返回如下包含错误信息的 JSON 报文：

```
{
  "cniVersion": "0.3.1",
  "code": <numeric-error-code>,
  "msg": <short-error-message>,
  "details": <long-error-message> (optional)
}
```

错误码包括如下内容。

- ◎ CNI 版本不匹配。
- ◎ 网络配置中存在不支持的字段，详细信息应在 msg 中说明。

### 3. 在 Kubernetes 中使用网络插件

Kubernetes 目前支持两种网络插件的实现。

- ◎ CNI 插件：根据 CNI 规范实现其接口，以与插件提供者进行对接。
- ◎ kubenet 插件：使用 bridge 和 host-local CNI 插件实现了一个基本的 cbr0，目前还处于 Alpha 版本阶段。

为了在 Kubernetes 集群中使用网络插件，需要在 kubelet 服务的启动参数上设置下面两个参数。

- ◎ --network-plugin-dir：kubelet 启动时扫描网络插件的目录。
- ◎ --network-plugin：网络插件名称，对于 CNI 插件，设置为“cni”即可，无须关注 --network-plugin-dir 的路径。对于 kubenet 插件，设置为“kubenet”，目前仅实现了一个简单的 cbr0 Linux 网桥。

在设置--network-plugin="cni"时，kubelet 还需设置下面两个参数。

- ◎ --cni-conf-dir：CNI 插件的配置文件目录，默认为/etc/cni/net.d。该目录下的配置文件内容需要符合 CNI 规范。
- ◎ --cni-bin-dir：CNI 插件的可执行文件目录，默认为/opt/cni/bin。

目前已有多个开源项目支持以 CNI 网络插件的形式部署到 Kubernetes 集群中，进行 Pod 的网络设置和网络策略的设置，包括 Calico、Canal、Cilium、Contiv、Flannel、Romana、Weave Net 等。

### 3.7.7 Kubernetes 网络策略

为了实现细粒度的容器间网络访问隔离策略，Kubernetes 从 v1.3 版本开始，由 SIG-Network 小组主导研发了 Network Policy 机制，目前 API 版本为 extensions/v1beta1。Network Policy 的主要功能是对 Pod 间的网络通信进行限制和准入控制，设置方式为将 Pod 的 Label 作为查询条件，设置允许访问或禁止访问的客户端 Pod 列表。目前查询条件可以作用于 Pod 和 Namespace 级别。

为了使用 Network Policy，Kubernetes 引入一个新的资源对象“NetworkPolicy”，供用户设置 Pod 间网络访问的策略。但仅定义一个网络策略是无法完成实际的网络隔离的，还需要一个策略控制器(policy controller)进行策略的实现。策略控制器由第三方网络组件提供，目前 Calico、Romana、Weave、OpenShift、OpenContrail 等开源项目均支持网络策略的实现。

Network Policy 的工作原理如图 3.36 所示，policy controller 需要实现一个 API Listener，监听用户设置的 NetworkPolicy 定义，并将网络访问规则通过各 Node 的 Agent 进行实际设置(Agent 则需要 CNI 网络插件实现)。

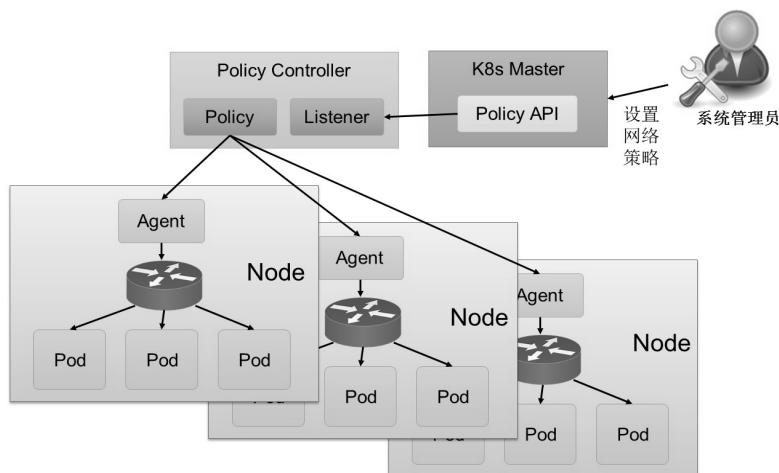


图 3.36 Network Policy 的工作原理

网络策略的设置包括网络隔离（禁止访问）和允许访问（白名单）两种方式，网络隔离需要在设置其他允许访问列表之前进行启用。

#### 1) 网络隔离

网络隔离作用于 Namespace 级别。为一个 Namespace 设置了网络隔离后，该规则将作用于属于这个 Namespace 的所有 Pod，即拒绝所有 Pod 之间的网络访问。当前仅支持“进入”方向的网络访问（Ingress）。

目前支持的网络隔离策略仅有 DefaultDeny，即到该 Namespace 中 Pod 的网络访问都被拒绝。

设置 DefaultDeny 网络策略方式为给 Namespace default 设置一个 annotation:

```
kind: Namespace
apiVersion: v1
metadata:
  name: default
  annotations:
    net.beta.kubernetes.io/network-policy: |
      {
        "ingress": {
          "isolation": "DefaultDeny"
        }
      }
}
```

也可以使用 kubectl 命令行工具给一个 Namespace 设置 annotation:

```
# kubectl annotate ns <namespace_name>
"net.beta.kubernetes.io/network-policy={\"ingress\": {\"isolation\":
\"DefaultDeny\"}}"
```

## 2) 准入白名单设置

在设置了默认的网络隔离策略之后，系统管理员还需对允许访问 Pod 的策略进行进一步设置，这就要通过 NetworkPolicy 对象来完成了，以下面的定义为例：

```
apiVersion: extensions/v1beta1
kind: NetworkPolicy
metadata:
  name: test-network-policy
  namespace: default
spec:
  podSelector:
    matchLabels:
      role: db
  ingress:
    - from:
      - namespaceSelector:
          matchLabels:
            project: myproject
      - podSelector:
          matchLabels:
            role: frontend
  ports:
    - protocol: tcp
      port: 6379
```

主要参数如下。

- ◎ **podSelector**: 用于定义后面 Ingress 策略将作用的 Pod 范围，即网络访问的目标 Pod 范围。

- ◎ ingress: 定义可以访问目标 Pod 的白名单策略, 满足 from 条件的客户端才能访问 ports 定义的目标 Pod 端口号。
  - from: 对符合条件的客户端 Pod 进行网络放行, 可以基于客户端 Pod 的 Label 或客户端 Pod 所在的 Namespace 的 Label 进行设置。
  - ports: 可访问的目标 Pod 监听的端口号。

通过这个 NetworkPolicy 定义, 管理员实际上设置了如下网络策略。

- ◎ 该策略作用于 Namespace default 中含有“role=db”Label 的全部 Pod(作为服务提供者)。
- ◎ 包含 “role=frontend” Label 的客户端 Pod 允许访问。
- ◎ 属于包含 “project=myproject” Label 的 Namespace 的客户端 Pod 允许访问。

下例的 NetworkPolicy 设置允许任何客户端访问:

```
kind: NetworkPolicy
apiVersion: extensions/v1beta1
metadata:
  name: allow-all
spec:
  podSelector:
    ingress:
      - {}
```

### 3.7.8 开源的网络组件

Kubernetes 的网络模型假定了所有 Pod 都在一个可以直接连通的扁平的网络空间中。这在 GCE 里面是现成的网络模型, Kubernetes 假定这个网络已经存在。而在私有云里搭建 Kubernetes 集群, 就不能假定这种网络已经存在了。我们需要自己实现这个网络假设, 将不同节点上的 Docker 容器之间的互相访问先打通, 然后运行 Kubernetes。

目前已经有多个开源组件支持容器网络模型。本节介绍几个常见的网络组件及其安装配置方法, 包括 Flannel、Open vSwitch、直接路由和 Calico。

#### 1. Flannel

Flannel 之所以可以搭建 Kubernetes 依赖的底层网络, 是因为它能实现以下两点。

- (1) 它能协助 Kubernetes, 给每一个 Node 上的 Docker 容器分配互相不冲突的 IP 地址。
- (2) 它能在这些 IP 地址之间建立一个覆盖网络 (Overlay Network), 通过这个覆盖网络, 将数据包原封不动地传递到目标容器内。

现在, 通过图 3.37 来看看 Flannel 是如何实现这两点的。

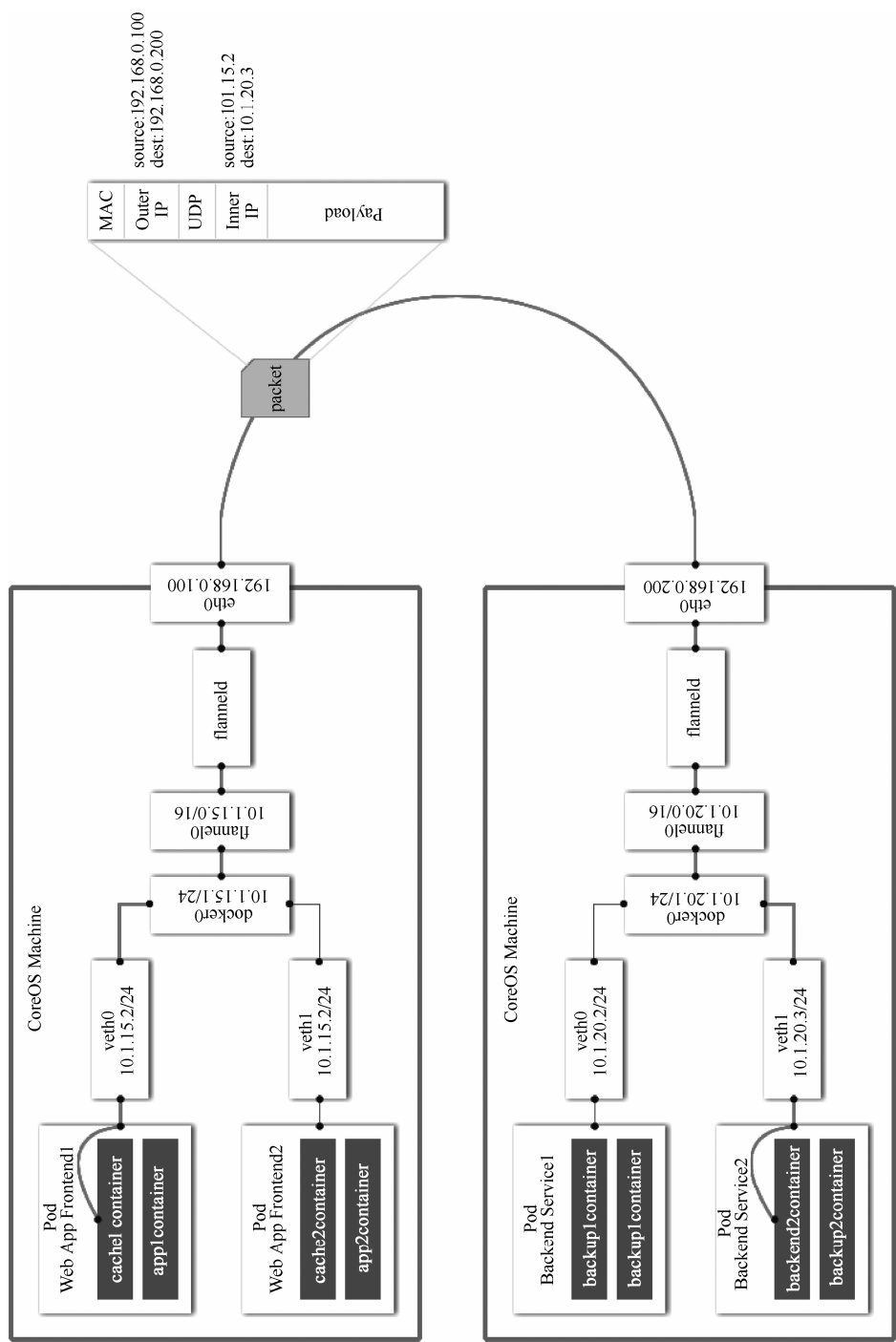


图 3.37 Flannel 架构图



可以看到, Flannel 首先创建了一个名为 flannel0 的网桥, 而且这个网桥的一端连接 docker0 网桥, 另一端连接一个叫作 flanneld 的服务进程。

flanneld 进程并不简单, 它首先上连 etcd, 利用 etcd 来管理可分配的 IP 地址段资源, 同时监控 etcd 中每个 Pod 的实际地址, 并在内存中建立了一个 Pod 节点路由表; 然后下连 docker0 和物理网络, 使用内存中的 Pod 节点路由表, 将 docker0 发给它的数据包包装起来, 利用物理网络的连接将数据包投递到目标 flanneld 上, 从而完成 Pod 到 Pod 之间的直接的地址通信。

Flannel 之间的底层通信协议的可选余地很多, 有 UDP、VxLan、AWS VPC 等多种方式, 只要能通到对端的 Flannel 就可以了。源 flanneld 加包, 目标 flanneld 解包, 最终 docker0 看到的的就是原始的数据, 非常透明, 根本感觉不到中间 Flannel 的存在。常用的是 UDP。

我们看一下 Flannel 是如何做到为不同 Node 上的 Pod 分配的 IP 不产生冲突的。其实想到 Flannel 使用了集中的 etcd 存储就很容易理解了。它每次分配的地址段都在同一个公共区域获取, 这样大家自然能够互相协调, 不产生冲突了。而且在 Flannel 分配好地址段后, 后面的事情是由 Docker 完成的, Flannel 通过修改 Docker 的启动参数将分配给它的地址段传递进去。

```
--bip=172.17.18.1/24
```

通过这些操作, Flannel 就控制了每个 Node 上的 docker0 地址段的地址, 也就保障了所有 Pod 的 IP 地址在同一个水平网络中且不产生冲突了。

Flannel 完美地实现了对 Kubernetes 网络的支持, 但是它引入了多个网络组件, 在网络通信时需要转到 flannel0 网络接口, 再转到用户态的 flanneld 程序, 到对端后还需要走这个过程的反过程, 所以也会引入一些网络的时延损耗。

另外, Flannel 模型默认使用了 UDP 作为底层传输协议, UDP 本身是非可靠协议, 虽然两端的 TCP 实现了可靠传输, 但在大流量、高并发应用场景下还需要反复测试, 确保没有问题。

Flannel 的安装和配置如下。

### 1) 安装 etcd

由于 Flannel 使用 etcd 作为数据库, 所以需要预先安装好 etcd, 此处不再赘述。

### 2) 安装 Flannel

需要在每台 Node 上都安装 Flannel。Flannel 软件的下载地址为 <https://github.com/coreos/flannel/releases>。将下载的压缩包 flannel-<version>-linux-amd64.tar.gz 解压, 把二进制文件 flanneld 和 mk-docker-opts.sh 复制到/usr/bin (或其他 PATH 环境变量中的目录) 中, 即可完成对 Flannel 的安装。

### 3) 配置 Flannel

此处以使用 systemd 系统为例对 flanneld 服务进行配置。编辑服务配置文件/usr/lib/systemd/system/flanneld.service:

```
[Unit]
Description=flanneld overlay address etcd agent
After=network.target
Before=docker.service

[Service]
Type=notify
EnvironmentFile=/etc/sysconfig/flanneld
ExecStart=/usr/bin/flanneld -etcd-endpoints=${FLANNEL_ETCD} $FLANNEL_OPTIONS

[Install]
RequiredBy=docker.service
WantedBy=multi-user.target
```

编辑配置文件/etc/sysconfig/flannel, 设置 etcd 的 URL 地址:

```
# flanneld configuration options

# etcd url location. Point this to the server where etcd runs
FLANNEL_ETCD="http://192.168.18.3:2379"

# etcd config key. This is the configuration key that flannel queries
# For address range assignment
FLANNEL_ETCD_KEY="/coreos.com/network"
```

在启动 flanneld 服务之前, 需要在 etcd 中添加一条网络配置记录, 这个配置将用于 flanneld 分配给每个 Docker 的虚拟 IP 地址段。

```
# etcdctl set /coreos.com/network/config '{ "Network": "10.1.0.0/16" }'
```

由于 Flannel 将覆盖 docker0 网桥, 所以如果 Docker 服务已启动, 则需要停止 Docker 服务。

### 4) 启动 flanneld 服务

```
# systemctl restart flanneld
```

### 5) 设置 docker0 网桥的 IP 地址

```
# mk-docker-opts.sh -i
# source /run/flannel/subnet.env
# ifconfig docker0 ${FLANNEL_SUBNET}
```

完成后确认网络接口 docker0 的 IP 地址属于 flannel0 的子网:

```
# ip addr
flannel0: flags=4305<UP,POINTOPOINT,RUNNING,NOARP,MULTICAST> mtu 1472
    inet 10.1.10.0 netmask 255.255.0.0 destination 10.1.10.0
```

```
docker0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500
    inet 10.1.10.1 netmask 255.255.255.0 broadcast 10.1.10.255
```

## 6) 重新启动 Docker 服务

```
# systemctl restart docker
```

至此就完成了 Flannel 覆盖网络的设置。

使用 ping 命令验证各 Node 上 docker0 之间的相互访问。例如在 Node1(docker0 IP=10.1.10.1) 机器上 ping Node2 的 docker0 (docker0's IP=10.1.30.1)，通过 Flannel 能够成功连接到其他物理机的 Docker 网络：

```
$ ping 10.1.30.1
PING 10.1.30.1 (10.1.30.1) 56(84) bytes of data.
64 bytes from 10.1.30.1: icmp_seq=1 ttl=62 time=1.15 ms
64 bytes from 10.1.30.1: icmp_seq=2 ttl=62 time=1.16 ms
64 bytes from 10.1.30.1: icmp_seq=3 ttl=62 time=1.57 ms
```

我们也可以在 etcd 中查看到 Flannel 设置的 flannel0 地址与物理机 IP 地址的对应规则：

```
# etcdctl ls /coreos.com/network/subnets
/coreos.com/network/subnets/10.1.10.0-24
/coreos.com/network/subnets/10.1.20.0-24
/coreos.com/network/subnets/10.1.30.0-24

# etcdctl get /coreos.com/network/subnets/10.1.10.0-24
{"PublicIP": "192.168.1.129"}
# etcdctl get /coreos.com/network/subnets/10.1.20.0-24
{"PublicIP": "192.168.1.130"}
# etcdctl get /coreos.com/network/subnets/10.1.30.0-24
{"PublicIP": "192.168.1.131"}
```

## 2. Open vSwitch

在了解了 Flannel 后，我们再看看 Open vSwitch 是怎么解决上述两个问题的。

Open vSwitch 是一个开源的虚拟交换机软件，有点儿像 Linux 中的 bridge，但是功能要复杂得多。Open vSwitch 的网桥可以直接建立多种通信通道（隧道），例如 Open vSwitch with GRE/VxLAN。这些通道的建立可以很容易地通过 OVS 的配置命令实现。在 Kubernetes、Docker 场景下，我们主要是建立 L3 到 L3 的隧道。举个例子来看看 Open vSwitch with GRE/VxLAN 的网络架构，如图 3.38 所示。

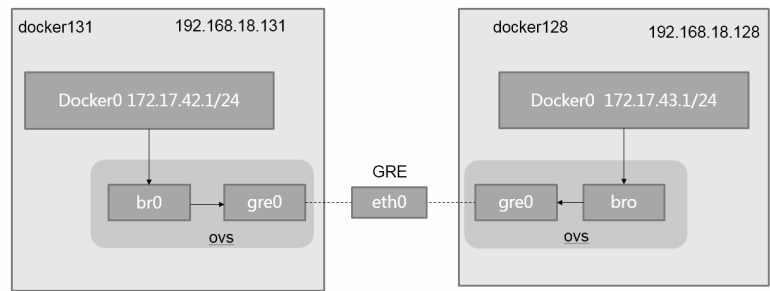


图 3.38 Open vSwitch with GRE/VxLAN 的网络架构

首先,为了避免 Docker 创建的 docker0 地址产生冲突(因为 Docker Daemon 启动且给 docker0 选择子网地址时只有几个备选列表,很容易产生冲突),我们可以将 docker0 网桥删除,手动建立一个 Linux 网桥,然后手动给这个网桥配置 IP 地址范围。

其次,建立 Open vSwitch 的网桥 ovs,然后使用 ovs-vsctl 命令给 ovs 网桥增加 gre 端口,添加 gre 端口时要将目标连接的 NodeIP 地址设置为对端的 IP 地址。对每一个对端 IP 地址都需要这么操作(对于大型集群网络,这可是个体力活,要做自动化脚本来完成)。

最后将 ovs 的网桥作为网络接口,加入 Docker 的网桥上(docker0 或者自己手工建立的新网桥)。

重启 ovs 网桥和 Docker 的网桥,并添加一个 Docker 的地址段到 Docker 网桥的路由规则项,就可以将两个容器的网络连接起来了。

1) 网络通信过程

当容器内的应用访问另一个容器的地址时,数据包会通过容器内的默认路由发送给 docker0 网桥。ovs 的网桥是作为 docker0 网桥的端口存在的,它会将数据发送给 ovs 网桥。ovs 网络已经通过配置建立了和其他 ovs 网桥的 GRE/VxLAN 隧道,自然能将数据送达对端的 Node,并送往 docker0 及 Pod。

通过新增的路由项,使得 Node 节点本身的应用的数据也路由到 docker0 网桥上,和刚才的通信过程一样,自然也可以访问其他 Node 上的 Pod。

2) OVS with GRE/VxLAN 组网方式的特点

OVS 的优势是,作为开源虚拟交换机软件,它相对比较成熟和稳定,而且支持各类网络隧道协议,经过了 OpenStack 等项目的考验。

另一方面,在前面介绍 Flannel 时可知,Flannel 除了支持建立覆盖网络(Overlay Network),保证 Pod 到 Pod 的无缝通信,还和 Kubernetes、Docker 架构体系结合紧密。Flannel 能够感知 Kubernetes 的 Service,动态维护自己的路由表,还通过 etcd 来协助 Docker 对整个 Kubernetes

集群中 docker0 的子网地址分配。而我们在使用 OVS 时，很多事情就需要手工完成了。

无论是 OVS 还是 Flannel，通过覆盖网络提供的 Pod 到 Pod 通信都会引入一些额外的通信开销，如果是对网络依赖特别重的应用，则需要评估对业务的影响。

Open vSwitch 的安装和配置如下。

以两个 Node 为例，目标网络拓扑如图 3.39 所示。

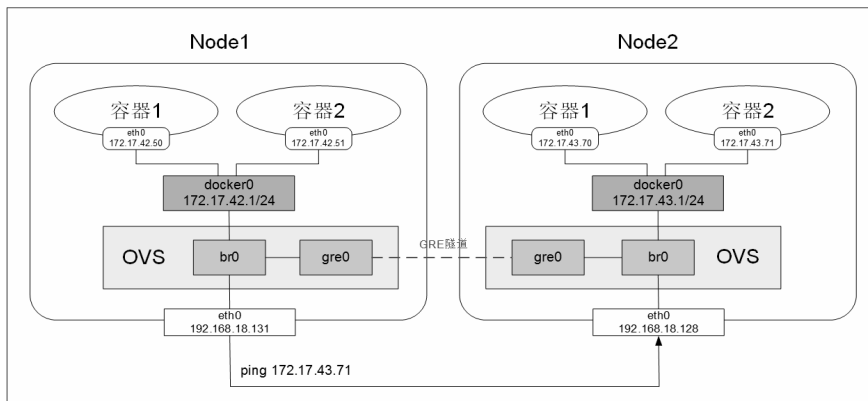


图 3.39 目标网络拓扑

首先，确保节点 192.168.18.128 的 Docker0 采用了 172.17.43.0/24 网段，而 192.168.18.131 的 Docker0 采用了 172.17.42.0/24 网段，对应的参数为 docker daemon 的启动参数--bip 设置的值。

### 1) 在两个 Node 上安装 ovs

```
# yum install openvswitch-2.4.0-1.x86_64.rpm
```

禁止 selinux，配置后重启 Linux：

```
# vi /etc/selinux/config
SELINUX=disabled
```

查看 Open vSwitch 的服务状态，应该启动 ovsdb-server 与 ovs-vswitchd 两个进程。

```
# service openvswitch status
ovsdb-server is running with pid 2429
ovs-vswitchd is running with pid 2439
```

查看 Open vSwitch 的相关日志，确认没有异常：

```
# more /var/log/messages |grep openv
Nov 2 03:12:52 docker128 openvswitch: Starting ovsdb-server [ OK ]
Nov 2 03:12:52 docker128 openvswitch: Configuring Open vSwitch system IDs
[ OK ]
Nov 2 03:12:52 docker128 kernel: openvswitch: Open vSwitch switching datapath
Nov 2 03:12:52 docker128 openvswitch: Inserting openvswitch module [ OK ]
```

注意，上述操作需要在两个节点机器上分别执行完成。

## 2) 创建网桥和 GRE 隧道

接下来需要在每个 Node 上建立 ovs 的网桥 br0，然后在网桥上创建一个 GRE 隧道连接对端网桥，最后把 ovs 的网桥 br0 作为一个端口连接到 docker0 这个 Linux 网桥上（可以认为是交换机互联），这样一来，两个节点机器上的 docker0 网段就能互通了。

下面以节点机器 192.168.18.131 为例，具体的操作步骤如下。

(1) 创建 ovs 网桥：

```
# ovs-vsctl add-br br0
```

(2) 创建 GRE 隧道连接对端，remote\_ip 为对端 eth0 的网卡地址：

```
# ovs-vsctl add-port br0 gre1 -- set interface gre1 type=gre
option:remote_ip=192.168.18.128
```

(3) 添加 br0 到本地 docker0，使得容器流量通过 OVS 流经 tunnel：

```
# brctl addif docker0 br0
```

(4) 启动 br0 与 docker0 网桥：

```
# ip link set dev br0 up
# ip link set dev docker0 up
```

(5) 添加路由规则。由于 192.168.18.128 与 192.168.18.131 的 docker0 网段分别为 172.17.43.0/24 与 172.17.42.0/24，这两个网段的路由都需要经过本机的 docker0 网桥路由，其中一个 24 网段是通过 OVS 的 GRE 隧道到达对端的，因此需要在每个 Node 上添加通过 docker0 网桥转发的 172.17.0.0/16 段的路由规则：

```
# ip route add 172.17.0.0/16 dev docker0
```

(6) 清空 Docker 自带的 Iptables 规则及 Linux 的规则，后者存在拒绝 icmp 报文通过防火墙的规则：

```
# iptables -t nat -F; iptables -F
```

在 192.168.18.131 上完成上述步骤后，在 192.168.18.128 节点执行同样的操作，注意，GRE 隧道里的 IP 地址要改为对端节点（192.168.18.131）的 IP 地址。

配置完成后，192.168.18.131 的 IP 地址、docker0 的 IP 地址及路由等重要信息显示如下：

```
# ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
2: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP
```

```

qlen 1000
    link/ether 00:0c:29:55:5e:c3 brd ff:ff:ff:ff:ff:ff
    inet 192.168.18.131/24 brd 192.168.18.255 scope global dynamic eth0
        valid_lft 1369sec preferred_lft 1369sec
3: ovs-system: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN
    link/ether a6:15:c3:25:cf:33 brd ff:ff:ff:ff:ff:ff
4: br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue master docker0
state UNKNOWN
    link/ether 92:8d:d0:a4:ca:45 brd ff:ff:ff:ff:ff:ff
5: docker0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP
    link/ether 02:42:44:8d:62:11 brd ff:ff:ff:ff:ff:ff
    inet 172.17.42.1/24 scope global docker0
        valid_lft forever preferred_lft forever

```

同样，192.168.18.128 节点的重要信息如下：

```

# ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
2: eth0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP
qlen 1000
    link/ether 00:0c:29:e8:02:c7 brd ff:ff:ff:ff:ff:ff
    inet 192.168.18.128/24 brd 192.168.18.255 scope global dynamic eth0
        valid_lft 1356sec preferred_lft 1356sec
3: ovs-system: <BROADCAST,MULTICAST> mtu 1500 qdisc noop state DOWN
    link/ether fa:6c:89:a2:f2:01 brd ff:ff:ff:ff:ff:ff
4: br0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue master docker0
state UNKNOWN
    link/ether ba:89:14:e0:7f:43 brd ff:ff:ff:ff:ff:ff
5: docker0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue state UP
    link/ether 02:42:63:a8:14:d5 brd ff:ff:ff:ff:ff:ff
    inet 172.17.43.1/24 scope global docker0
        valid_lft forever preferred_lft forever

```

### 3) 两个 Node 上容器之间的互通测试

首先，在 192.168.18.128 节点上 ping 192.168.18.131 上的 docker0 地址 172.17.42.1，验证网络的互通性：

```

# ping 172.17.42.1
PING 172.17.42.1 (172.17.42.1) 56(84) bytes of data.
64 bytes from 172.17.42.1: icmp_seq=1 ttl=64 time=1.57 ms
64 bytes from 172.17.42.1: icmp_seq=2 ttl=64 time=0.966 ms
64 bytes from 172.17.42.1: icmp_seq=3 ttl=64 time=1.01 ms
64 bytes from 172.17.42.1: icmp_seq=4 ttl=64 time=1.00 ms

```

```
64 bytes from 172.17.42.1: icmp_seq=5 ttl=64 time=1.22 ms
64 bytes from 172.17.42.1: icmp_seq=6 ttl=64 time=0.996 ms
```

下面我们通过 `tshark` 抓包工具来分析流量走向。首先，在 `192.168.18.128` 节点上监听 `br0` 上是否有 GRE 报文，执行下面的命令，我们发现 `br0` 上并没有 GRE 报文：

```
# tshark -i br0 -R ip proto GRE
tshark: -R without -2 is deprecated. For single-pass filtering use -Y.
Running as user "root" and group "root". This could be dangerous.
Capturing on 'br0'
^C
```

而在 `eth0` 上抓包，则发现了 GRE 封装的 ping 包报文通过，说明 GRE 是在承载网的物理网上完成的封包过程：

```
# tshark -i eth0 -R ip proto GRE
tshark: -R without -2 is deprecated. For single-pass filtering use -Y.
Running as user "root" and group "root". This could be dangerous.
Capturing on 'eth0'
 1  0.000000 172.17.43.1 -> 172.17.42.1  ICMP 136 Echo (ping) request
id=0x0970, seq=180/46080, ttl=64
 2  0.000892 172.17.42.1 -> 172.17.43.1  ICMP 136 Echo (ping) reply
id=0x0970, seq=180/46080, ttl=64 (request in 1)
 2  3  1.002014 172.17.43.1 -> 172.17.42.1  ICMP 136 Echo (ping) request
id=0x0970, seq=181/46336, ttl=64
 4  1.002916 172.17.42.1 -> 172.17.43.1  ICMP 136 Echo (ping) reply
id=0x0970, seq=181/46336, ttl=64 (request in 3)
 4  5  2.004101 172.17.43.1 -> 172.17.42.1  ICMP 136 Echo (ping) request
id=0x0970, seq=182/46592, ttl=64
```

至此，基于 OVS 的网络搭建成功，由于 GRE 是点对点的隧道通信方式，所以如果有多个 Node，则需要建立  $N \times (N-1)$  条 GRE 隧道，即所有 Node 组成一个网状的网络，实现了全网互通。

### 3. 直接路由

我们知道，`docker0` 网桥上的 IP 地址在 Node 网络上是不看到的。从一个 Node 到一个 Node 内的 `docker0` 是不通的，因为它不知道某个 IP 地址在哪里。如果能够让这些机器知道对端 `docker0` 地址在哪里，就可以让这些 `docker0` 互相通信了。这样所有 Node 上运行的 Pod 就可以互相通信了。

我们可以通过部署 MultiLayer Switch (MLS) 来实现这一点，在 MLS 中配置每个 `docker0` 子网地址到 Node 地址的路由项，通过 MLS 将 `docker0` 的 IP 寻址定向到对应的 Node 节点上。



另外，我们可以将这些 docker0 和 Node 的匹配关系配置在 Linux 操作系统的路由项中，这样通信发起的 Node 能够根据这些路由信息直接找到目标 Pod 所在的 Node，将数据传输过去。如图 3.40 所示。

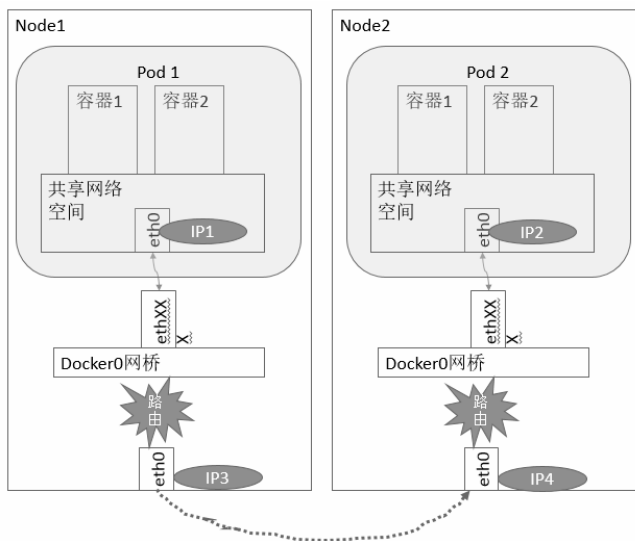


图 3.40 直接路由 Pod 到 Pod 通信

我们在每个 Node 的路由表中增加对方所有 docker0 的路由项。

例如 Pod1 所在 docker0 网桥的 IP 子网是 10.1.10.0，Node 的地址为 192.168.1.128；而 Pod2 所在 docker0 网桥的 IP 子网是 10.1.20.0，Node 的地址为 192.168.1.129。

在 Node1 上用 `route add` 命令增加一条到 Node2 上 docker0 的静态路由规则：

```
# route add -net 10.1.20.0 netmask 255.255.255.0 gw 192.168.1.129
```

同样，在 Node2 上增加一条到 Node1 上 docker0 的静态路由规则：

```
# route add -net 10.1.10.0 netmask 255.255.255.0 gw 192.168.1.128
```

在 Node1 上通过 `ping` 命令验证到 Node2 上 docker0 的网络连通性。这里 10.1.20.1 为 Node2 上 docker0 网桥自身的 IP 地址。

```
$ ping 10.1.20.1
PING 10.1.20.1 (10.1.20.1) 56(84) bytes of data.
64 bytes from 10.1.20.1: icmp_seq=1 ttl=62 time=1.15 ms
64 bytes from 10.1.20.1: icmp_seq=2 ttl=62 time=1.16 ms
64 bytes from 10.1.20.1: icmp_seq=3 ttl=62 time=1.57 ms
.....
```

可以看到，路由转发规则生效，Node1 可以直接访问 Node2 上的 docker0 网桥，进一步可

以访问属于 `docker0` 网段的容器应用了。

在大规模集群中，在每个 Node 上都需要配置到其他 `docker0/Node` 的路由项，这会带来很大的工作量；并且在新增机器时，对所有 Node 都需要修改配置；重启机器时，如果 `docker0` 的地址有变化，则也需要修改所有 Node 的配置，这显然是非常复杂的。

为了管理这些动态变化的 `docker0` 地址，动态地让其他 Node 都感知到它，还可以使用动态路由发现协议来同步这些变化。运行动态路由发现协议代理的 Node 时，会将本机 LOCAL 路由表的 IP 地址通过组播协议发布出去，同时监听其他 Node 的组播包。通过这样的信息交换，Node 上的路由规则都能够相互学习到。当然，路由发现协议本身还是很复杂的，感兴趣的话你可以查阅相关规范。在实现这些动态路由发现协议的开源软件中，常用的有 Quagga (<http://www.quagga.net>)、Zebra 等。下面简单介绍直接路由的操作过程。

(1) 首先手工分配 Docker bridge 的地址，保证它们在不同的网段是不重叠的。建议最好不用 Docker Daemon 自动创建的 `docker0`（因为我们不需要它的自动管理功能），而是单独建立一个 bridge，给它配置规划好的 IP 地址，然后使用 `--bridge=XX` 来指定网桥。

(2) 然后在每个节点上运行 Quagga。

完成这些操作后，我们很快就能得到一个 Pod 和 Pod 直接互相访问的环境了。由于路由发现能够被网络上的所有设备接收，所以如果网络上的路由器也能打开 RIP 协议选项，则能够学习到这些路由信息。通过这些路由器，我们甚至可以在非 Node 节点上使用 Pod 的 IP 地址直接访问 Node 上的 Pod 了。

除了在每台服务器上安装 Quagga 软件并启动，还可以使用 Quagga 容器来运行（例如 `index.alauda.cn/georce/router`）。在每台 Node 上下载该 Docker 镜像：

```
$ docker pull index.alauda.cn/georce/router
```

在运行 Quagga 容器之前，需要确保每个 Node 上 `docker0` 网桥的子网地址不能重叠，也不能与物理机所在的网络重叠，这需要网络管理员的仔细规划。

下面以 3 台 Node 为例，每台 Node 的 `docker0` 网桥的地址如下（前提是 Node 物理机的 IP 地址不是 10.1.X.X 地址段）：

```
Node 1: # ifconfig docker0 10.1.10.1/24
Node 2: # ifconfig docker0 10.1.20.1/24
Node 3: # ifconfig docker0 10.1.30.1/24
```

然后在每台 Node 上启动 Quagga 容器。需要说明的是，Quagga 需要以 `--privileged` 特权模式运行，并且指定 `--net=host`，表示直接使用物理机的网络：

```
$ docker run -itd --name=router --privileged --net=host index.alauda.cn/georce/router
```

启动成功后，各 Node 上的 Quagga 会相互学习来完成到其他机器的 docker0 路由规则的添加。

一段时间后，在 Node1 上使用 `route -n` 命令来查看路由表，可以看到 Quagga 自动添加了两条到 Node2 和到 Node3 上 docker0 的路由规则。

```
# route -n
Kernel IP routing table
Destination      Gateway          Genmask          Flags  Metric  Ref    Use Iface
0.0.0.0          192.168.1.128   0.0.0.0          UG      0        0      0 eth0
10.1.10.0        0.0.0.0         255.255.255.0    U        0        0      0 docker0
10.1.20.0        192.168.1.129   255.255.255.0    UG      20        0      0 eth0
10.1.30.0        192.168.1.130   255.255.255.0    UG      20        0      0 eth0
```

在 Node2 上查看路由表，可以看到自动添加了两条到 Node1 和 Node3 上 docker0 的路由规则。

```
# route -n
Kernel IP routing table
Destination      Gateway          Genmask          Flags  Metric  Ref    Use Iface
0.0.0.0          192.168.1.129   0.0.0.0          UG      0        0      0 eth0
10.1.20.0        0.0.0.0         255.255.255.0    U        0        0      0 docker0
10.1.10.0        192.168.1.128   255.255.255.0    UG      20        0      0 eth0
10.1.30.0        192.168.1.130   255.255.255.0    UG      20        0      0 eth0
```

至此，所有 Node 上的 docker0 都可以互联互通了。

当然，聪明的你还会有新的疑问：这样做的话，由于每个 Pod 的地址都会被路由发现协议广播出去，会不会存在路由表过大的情况？实际上，路由表通常都会有高速缓存，查找速度会很快，不会对性能产生太大的影响。当然，如果你的集群容量在数千台 Node 以上，则仍然需要测试和评估路由表的效率问题。

## 4. Calico 容器网络和网络策略实战

本节以 Calico 为例讲解在 Kubernetes 中 CNI 插件和网络策略的原理和应用。

### 1) Calico 简介

Calico 是一个基于 BGP 的纯三层的网络方案，与 OpenStack、Kubernetes、AWS、GCE 等云平台都能够良好地集成。Calico 在每个计算节点利用 Linux Kernel 实现了一个高效的 vRouter 来负责数据转发。每个 vRouter 通过 BGP1 协议把在本节点上运行的容器的路由信息向整个 Calico 网络广播，并自动设置到达其他节点的路由转发规则。Calico 保证所有容器之间的数据流量都是通过 IP 路由的方式完成互联互通的。Calico 节点组网可以直接利用数据中心的网络结构（L2 或者 L3），不需要额外的 NAT、隧道或者 Overlay Network，没有额外的封包解包，能

够节约 CPU 运算，提高网络效率，如图 3.41 所示。

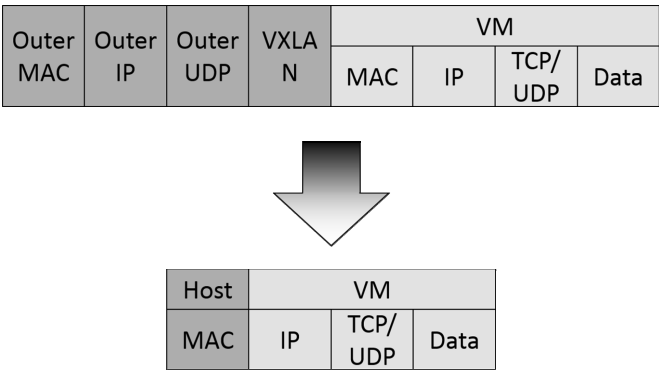


图 3.41 Calico 不使用额外的封包解包

Calico 在小规模集群中可以直接互联，在大规模集群中可以通过额外的 BGP route reflector 来完成，如图 3.42 所示。

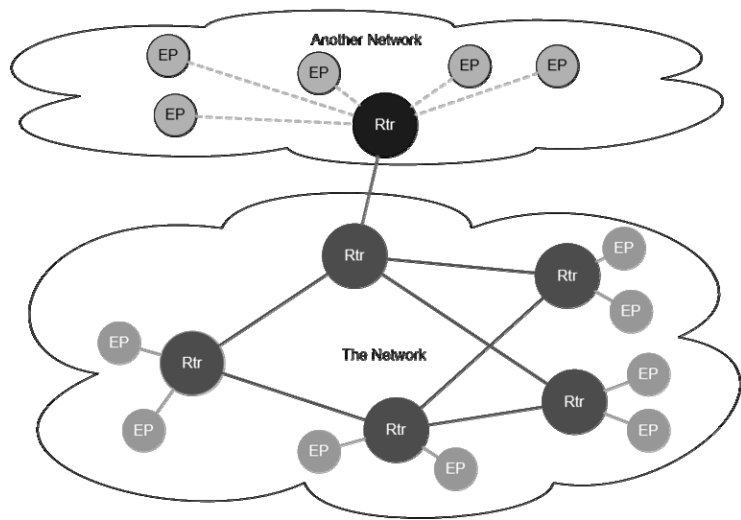


图 3.42 通过 BGP route reflector 连接大规模网络

此外，Calico 基于 Iptables 还提供了丰富的网络策略，实现了 Kubernetes 的 Network Policy 策略，提供容器间网络可达性限制的功能。

Calico 的系统架构如图 3.43 所示。

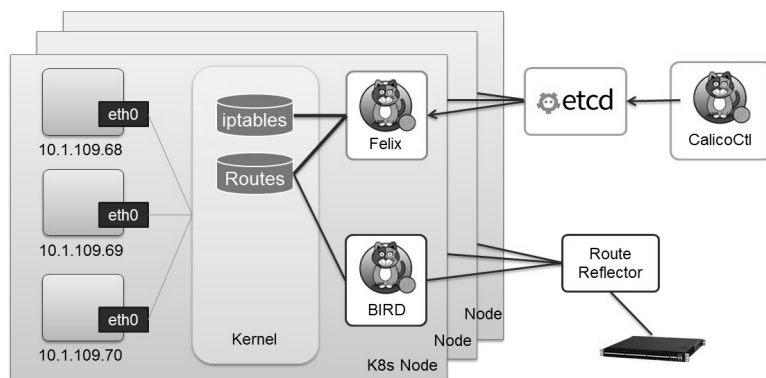


图 3.43 Calico 的系统架构

Calico 的主要组件如下。

- ◎ **Felix:** Calico Agent，运行在每台 Node 上，负责为容器设置网络资源（IP 地址、路由规则、Iptables 规则等），保证跨主机容器网络互通。
- ◎ **etcd:** Calico 使用的后端存储。
- ◎ **BGP Client (BIRD):** 负责把 Felix 在各 Node 上设置的路由信息通过 BGP 协议广播到 Calico 网络。
- ◎ **BGP Route Reflector (BIRD):** 通过一个或者多个 BGP Route Reflector 来完成大规模集群的分级路由分发。
- ◎ **calicoctl:** Calico 命令行管理工具。

## 2) 部署 Calico 服务

在 Kubernetes 中部署 Calico 的主要步骤如下。

(1) 修改 Kubernetes 服务的启动参数，并重启服务。

- ◎ 设置 Master 上 kube-apiserver 服务的启动参数: `--allow-privileged=true` (因为 calico-node 需要以特权模式运行在各 Node 上)。
- ◎ 设置各 Node 上 kubelet 服务的启动参数: `--network-plugin=cni` (使用 CNI 网络插件)。

本例中的 Kubernetes 集群包括两台 Node: k8s-node-1 (IP 地址为 192.168.18.3) 和 k8s-node-2 (IP 地址为 192.168.18.4)。

(2) 创建 Calico 服务，主要包括 calico-node 和 calico policy controller。需要创建的资源对象如下。

- ◎ 创建 ConfigMap calico-config，包含 Calico 所需的配置参数。

- ◎ 创建 Secret calico-etcd-secrets，用于使用 TLS 方式连接 etcd。
- ◎ 在每个 Node 上运行 calico/node 容器，部署为 DaemonSet。
- ◎ 在每个 Node 上安装 Calico CNI 二进制文件和网络配置参数（由 install-cni 容器完成）。
- ◎ 部署一个名为 calico/kube-policy-controller 的 Deployment，以对接 Kubernetes 集群中为 Pod 设置的 Network Policy。

从 Calico 官网下载 Calico 的 yaml 配置文件，下载地址为 <http://docs.projectcalico.org/v2.1/getting-started/kubernetes/installation/hosted/calico.yaml>，该配置文件包括了启动 Calico 所需的全部资源对象的定义，下面对它们逐个进行说明。

(1) Calico 所需的配置，以 ConfigMap 对象进行创建：

```
kind: ConfigMap
apiVersion: v1
metadata:
  name: calico-config
  namespace: kube-system
data:
  # Configure this with the location of your etcd cluster.
  etcd_endpoints: "http://192.168.18.3:2379"

  # Configure the Calico backend to use.
  calico_backend: "bird"

  # The CNI network configuration to install on each node.
  cni_network_config: |-
    {
      "name": "k8s-pod-network",
      "type": "calico",
      "etcd_endpoints": "__ETCD_ENDPOINTS__",
      "etcd_key_file": "__ETCD_KEY_FILE__",
      "etcd_cert_file": "__ETCD_CERT_FILE__",
      "etcd_ca_cert_file": "__ETCD_CA_CERT_FILE__",
      "log_level": "info",
      "ipam": {
        "type": "calico-ipam"
      },
      "policy": {
        "type": "k8s",
        "k8s_api_root":
"https://__KUBERNETES_SERVICE_HOST__:__KUBERNETES_SERVICE_PORT__",
        "k8s_auth_token": "__SERVICEACCOUNT_TOKEN__"
      },
      "kubernetes": {
```

```

    "kubeconfig": "__KUBECONFIG_FILEPATH__"
  }
}

# If you're using TLS enabled etcd uncomment the following.
# You must also populate the Secret below with these files.
etcd_ca: "" # "/calico-secrets/etcd-ca"
etcd_cert: "" # "/calico-secrets/etcd-cert"
etcd_key: "" # "/calico-secrets/etcd-key"

```

主要参数如下。

- ◎ **etcd\_endpoints**: Calico 使用 etcd 来保存网络拓扑和状态，该参数指定 etcd 的地址，可以使用 Kubernetes Master 所用的 etcd，也可以另外搭建。
- ◎ **calico\_backend**: Calico 的后端，默认为 bird。
- ◎ **cni\_network\_config**: 符合 CNI 规范的网络配置。其中 **type=calico** 表示 kubelet 将从 `/opt/cni/bin` 目录下搜索名为“calico”的可执行文件，并调用它完成容器网络的设置。  
**ipam** 中 **type=calico-ipam** 表示 kubelet 将在 `/opt/cni/bin` 目录下搜索名为“calico-ipam”的可执行文件，用于完成容器 IP 地址的分配。

etcd 如果配置了 TLS 安全认证，则还需指定相应的 **ca**、**cert**、**key** 等文件。

(2) 访问 etcd 所需的 secret，对于无 TLS 的 etcd 服务，将 **data** 设置为空即可：

```

apiVersion: v1
kind: Secret
type: Opaque
metadata:
  name: calico-etcd-secrets
  namespace: kube-system
data:
  # Populate the following files with etcd TLS configuration if desired, but leave
blank if
  # not using TLS for etcd.
  # This self-hosted install expects three files with the following names. The
values
  # should be base64 encoded strings of the entire contents of each file.
  # etcd-key: null
  # etcd-cert: null
  # etcd-ca: null

```

(3) calico-node, 以 Daemonset 形式在每台 Node 上运行一个 calico-node 服务和一个 install-cni 服务：

```

kind: DaemonSet
apiVersion: extensions/v1beta1

```

```
metadata:
  name: calico-node
  namespace: kube-system
  labels:
    k8s-app: calico-node
spec:
  selector:
    matchLabels:
      k8s-app: calico-node
  template:
    metadata:
      labels:
        k8s-app: calico-node
    annotations:
      scheduler.alpha.kubernetes.io/critical-pod: ''
      scheduler.alpha.kubernetes.io/tolerations: |
        [{"key": "dedicated", "value": "master", "effect": "NoSchedule" },
        {"key": "CriticalAddonsOnly", "operator": "Exists"}]
  spec:
    hostNetwork: true
    containers:
      # Runs calico/node container on each Kubernetes node. This
      # container programs network policy and routes on each
      # host.
      - name: calico-node
        image: quay.io/calico/node:v1.2.1
        env:
          # The location of the Calico etcd cluster.
          - name: ETCD_ENDPOINTS
            valueFrom:
              configMapKeyRef:
                name: calico-config
                key: etcd_endpoints
          # Choose the backend to use.
          - name: CALICO_NETWORKING_BACKEND
            valueFrom:
              configMapKeyRef:
                name: calico-config
                key: calico_backend
          # Disable file logging so `kubectl logs` works.
          - name: CALICO_DISABLE_FILE_LOGGING
            value: "true"
          # Set Felix endpoint to host default action to ACCEPT.
          - name: FELIX_DEFAULTENDPOINTTOHOSTACTION
            value: "ACCEPT"
          # Configure the IP Pool from which Pod IPs will be chosen.
          - name: CALICO_IPV4POOL_CIDR
```



```

    value: "10.1.0.0/16"
  - name: CALICO_IPV4POOL_IPIP
    value: "always"
  # IP Autodetection methods
  - name: IP_AUTODETECTION_METHOD
    value: "interface=ens.*"
  - name: IP6_AUTODETECTION_METHOD
    value: "interface=ens.*"
  # Disable IPv6 on Kubernetes.
  - name: FELIX_IPV6SUPPORT
    value: "false"
  # Set Felix logging to "info"
  - name: FELIX_LOGSEVERITYSCREEN
    value: "info"
  # Location of the CA certificate for etcd.
  - name: ETCD_CA_CERT_FILE
    valueFrom:
      configMapKeyRef:
        name: calico-config
        key: etcd_ca
  # Location of the client key for etcd.
  - name: ETCD_KEY_FILE
    valueFrom:
      configMapKeyRef:
        name: calico-config
        key: etcd_key
  # Location of the client certificate for etcd.
  - name: ETCD_CERT_FILE
    valueFrom:
      configMapKeyRef:
        name: calico-config
        key: etcd_cert
  # Auto-detect the BGP IP address.
  - name: IP
    value: ""
securityContext:
  privileged: true
resources:
  requests:
    cpu: 250m
volumeMounts:
  - mountPath: /lib/modules
    name: lib-modules
    readOnly: true
  - mountPath: /var/run/calico
    name: var-run-calico
    readOnly: false

```

```
- mountPath: /calico-secrets
  name: etcd-certs
# This container installs the Calico CNI binaries
# and CNI network config file on each node.
- name: install-cni
  image: quay.io/calico/cni:v1.8.3
  command: ["/install-cni.sh"]
  env:
    # The location of the Calico etcd cluster.
    - name: ETCD_ENDPOINTS
      valueFrom:
        configMapKeyRef:
          name: calico-config
          key: etcd_endpoints
    # The CNI network config to install on each node.
    - name: CNI_NETWORK_CONFIG
      valueFrom:
        configMapKeyRef:
          name: calico-config
          key: cni_network_config
  volumeMounts:
    - mountPath: /host/opt/cni/bin
      name: cni-bin-dir
    - mountPath: /host/etc/cni/net.d
      name: cni-net-dir
    - mountPath: /calico-secrets
      name: etcd-certs
volumes:
  # Used by calico/node.
  - name: lib-modules
    hostPath:
      path: /lib/modules
  - name: var-run-calico
    hostPath:
      path: /var/run/calico
  # Used to install CNI.
  - name: cni-bin-dir
    hostPath:
      path: /opt/cni/bin
  - name: cni-net-dir
    hostPath:
      path: /etc/cni/net.d
  # Mount in the etcd TLS secrets.
  - name: etcd-certs
    secret:
      secretName: calico-etcd-secrets
```

该 Pod 中包括如下两个容器。

- ◎ **calico-node**: Calico 服务程序，用于设置 Pod 的网络资源，保证 Pod 的网络与各 Node 互联互通。它还需要以 **hostNetwork** 模式运行，直接使用宿主机网络。
- ◎ **install-cni**: 在各 Node 上安装 CNI 二进制文件到 `/opt/cni/bin` 目录下，并安装相应的网络配置文件到 `/etc/cni/net.d` 目录下。

calico-node 服务的主要参数如下。

- ◎ **CALICO\_IPV4POOL\_CIDR**: Calico IPAM 的 IP 地址池，Pod 的 IP 地址将从该池中进行分配。
- ◎ **CALICO\_IPV4POOL\_IPIP**: 是否启用 IPIP 模式。启用 IPIP 模式时，Calico 将在 Node 上创建一个名为 “tunl0” 的虚拟隧道。
- ◎ **IP\_AUTODETECTION\_METHOD**: 获取 Node IP 地址的方式，默认使用第 1 个网络接口的 IP 地址，对于安装了多块网卡的 Node，可以使用正则表达式选择正确的网卡，例如 “`interface=ens.*`” 表示选择名称以 “ens” 开头的网卡的 IP 地址。
- ◎ **FELIX\_IPV6SUPPORT**: 是否启用 IPV6。
- ◎ **FELIX\_LOGSEVERITYSCREEN**: 日志级别。

IP Pool 可以使用两种模式: BGP 或 IPIP。使用 IPIP 模式时，设置 **CALICO\_IPV4POOL\_IPIP=“always”**，不使用 IPIP 模式时，设置 **CALICO\_IPV4POOL\_IPIP=“off”**，此时将使用 BGP 模式。

IPIP 是一种将各 Node 的路由之间做一个 tunnel，再把两个网络连接起来的模式，如图 3.44 所示。启用 IPIP 模式时，Calico 将在各 Node 上创建一个名为 “tunl0” 的虚拟网络接口。

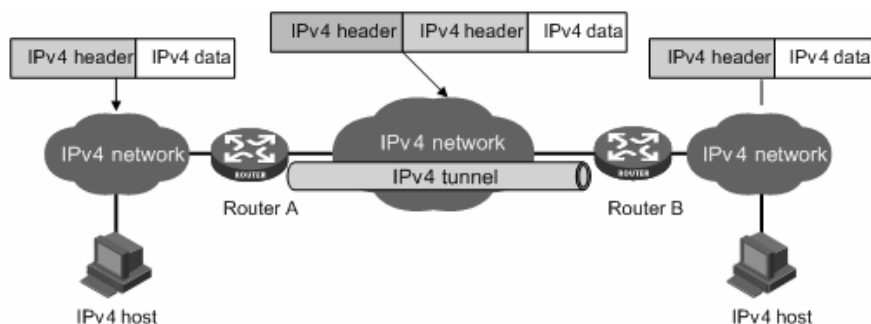


图 3.44 IPIP 模式

BGP 模式则直接使用物理机作为虚拟路由器 (vRouter)，不再创建额外的 tunnel。

(4) **calico-policy-controller** 容器，用于对接 Kubernetes 集群中为 Pod 设置的 Network Policy:

```
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: calico-policy-controller
  namespace: kube-system
  labels:
    k8s-app: calico-policy
  annotations:
    scheduler.alpha.kubernetes.io/critical-pod: ''
    scheduler.alpha.kubernetes.io/tolerations: |
      [{"key": "dedicated", "value": "master", "effect": "NoSchedule" },
      {"key": "CriticalAddonsOnly", "operator": "Exists"}]
spec:
  # The policy controller can only have a single active instance.
  replicas: 1
  strategy:
    type: Recreate
  template:
    metadata:
      name: calico-policy-controller
      namespace: kube-system
      labels:
        k8s-app: calico-policy
    spec:
      # The policy controller must run in the host network namespace so that
      # it isn't governed by policy that would prevent it from working.
      hostNetwork: true
      containers:
        - name: calico-policy-controller
          image: quay.io/calico/kube-policy-controller:v0.5.4
          env:
            # The location of the Calico etcd cluster.
            - name: ETCD_ENDPOINTS
              valueFrom:
                configMapKeyRef:
                  name: calico-config
                  key: etcd_endpoints
            # Location of the CA certificate for etcd.
            - name: ETCD_CA_CERT_FILE
              valueFrom:
                configMapKeyRef:
                  name: calico-config
                  key: etcd_ca
            # Location of the client key for etcd.
            - name: ETCD_KEY_FILE
              valueFrom:
                configMapKeyRef:
```

```

        name: calico-config
        key: etcd_key
    # Location of the client certificate for etcd.
    - name: ETCD_CERT_FILE
      valueFrom:
        configMapKeyRef:
          name: calico-config
          key: etcd_cert
    # The location of the Kubernetes API. Use the default Kubernetes
    # service for API access.
    - name: K8S_API
      value: "https://kubernetes.default:443"
    # Since we're running in the host namespace and might not have KubeDNS
    # access, configure the container's /etc/hosts to resolve
    # kubernetes.default to the correct service clusterIP.
    - name: CONFIGURE_ETC_HOSTS
      value: "true"
  volumeMounts:
    # Mount in the etcd TLS secrets.
    - mountPath: /calico-secrets
      name: etcd-certs
  volumes:
    # Mount in the etcd TLS secrets.
    - name: etcd-certs
      secret:
        secretName: calico-etcd-secrets

```

用户在 Kubernetes 集群中设置了 Pod 的 Network Policy 之后, calico-policy-controller 就会自动通知各 Node 上的 calico-node 服务, 在宿主机上设置相应的 Iptables 规则, 完成 Pod 间网络访问策略的设置。

修改好相应的参数后, 创建 Calico 的各资源对象:

```

# kubectl create -f calico.yaml
configmap "calico-config" created
secret "calico-etcd-secrets" created
daemonset "calico-node" created
deployment "calico-policy-controller" created

```

确保 Calico 各服务正确运行:

```

# kubectl get pods --namespace=kube-system -o wide

```

NAME	READY	STATUS	RESTARTS	AGE	IP
NODE					
calico-node-pgwqr	2/2	Running	0	1m	192.168.18.4
k8s-node-2					
calico-node-t3ntq	2/2	Running	0	1m	192.168.18.3
k8s-node-1					

```
calico-policy-controller-1838634297-cfddl 1/1 Running 0 2m
192.168.18.3 k8s-node-1
```

calico-node 在正常运行之后,会根据 CNI 规范,在/etc/cni/net.d/目录下生成如下文件和目录,并在/opt/cni/bin/目录下安装二进制文件 calico 和 calico-ipam,供 kubelet 调用。

- ◎ 10-calico.conf: 符合 CNI 规范的网络配置,其中 type=calico 表示该插件的二进制文件名为 calico。
- ◎ calico-kubeconfig: Calico 所需的 kubeconfig 文件。
- ◎ calico-tls 目录: 以 TLS 方式连接 etcd 的相关文件。

查看 k8s-node-1 服务器的网络接口设置,可以看到一个新的名为“tunl0”的接口,并设置了网络地址为 10.1.109.64/32:

```
# ip addr show
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN qlen 1
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
        valid_lft forever preferred_lft forever
2: ens33: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP
   qlen 1000
    link/ether 00:0c:29:1b:c5:fc brd ff:ff:ff:ff:ff:ff
    inet 192.168.18.3/24 brd 192.168.18.255 scope global ens33
        valid_lft forever preferred_lft forever
    inet6 fe80::20c:29ff:fe1b:c5fc/64 scope link
        valid_lft forever preferred_lft forever
3: docker0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc noqueue state DOWN
    link/ether 02:42:46:ad:a4:38 brd ff:ff:ff:ff:ff:ff
    inet 172.17.1.1/24 scope global docker0
        valid_lft forever preferred_lft forever
4: tunl0@NONE: <NOARP,UP,LOWER_UP> mtu 1440 qdisc noqueue state UNKNOWN qlen 1
    link/ipip 0.0.0.0 brd 0.0.0.0
    inet 10.1.109.64/32 scope global tunl0
        valid_lft forever preferred_lft forever
```

查看 k8s-node-2 服务器的网络接口设置,同样可以看到一个新的名为“tunl0”的接口,网络地址为 10.1.140.64/32:

```
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN qlen 1
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
        valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
        valid_lft forever preferred_lft forever
```

```

2: ens33: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc pfifo_fast state UP
qlen 1000
    link/ether 00:0c:29:93:71:9e brd ff:ff:ff:ff:ff:ff
    inet 192.168.18.4/24 brd 192.168.18.255 scope global ens33
        valid_lft forever preferred_lft forever
    inet6 fe80::20c:29ff:fe93:719e/64 scope link
        valid_lft forever preferred_lft forever
3: docker0: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc noqueue state DOWN
    link/ether 02:42:d9:08:8e:93 brd ff:ff:ff:ff:ff:ff
    inet 172.17.2.1/24 scope global docker0
        valid_lft forever preferred_lft forever
4: tunl0@NONE: <NOARP,UP,LOWER_UP> mtu 1440 qdisc noqueue state UNKNOWN qlen 1
    link/ipip 0.0.0.0 brd 0.0.0.0
    inet 10.1.140.64/32 scope global tunl0
        valid_lft forever preferred_lft forever

```

这两个子网都是从 calico-node 设置的 IP 地址池(CALICO\_IPV4POOL\_CIDR="10.1.0.0/16")中进行分配的。同时, docker0 对于 Kubernetes 设置 Pod 的 IP 地址将不再起作用。

查看两台主机的路由表。首先,查看 k8s-node-1 服务器的路由表,可以看到一条到 k8s-node-2 的私网 10.1.140.64 的路由转发规则:

```

# ip route
default via 192.168.18.2 dev ens33
blackhole 10.1.109.64/26 proto bird
10.1.140.64/26 via 192.168.18.4 dev tunl0 proto bird onlink
172.17.1.0/24 dev docker0 proto kernel scope link src 172.17.1.1
192.168.18.0/24 dev ens33 proto kernel scope link src 192.168.18.3 metric 100

```

然后,查看 k8s-node-2 服务器的路由表,可以看到一条到 k8s-node-1 的私网 10.1.109.64/26 的路由转发规则:

```

# ip route
default via 192.168.18.2 dev ens33
blackhole 10.1.140.64/26 proto bird
10.1.109.64/26 via 192.168.18.3 dev tunl0 proto bird onlink
172.17.2.0/24 dev docker0 proto kernel scope link src 172.17.2.1
192.168.18.0/24 dev ens33 proto kernel scope link src 192.168.18.4 metric 100

```

这样,通过 Calico 就完成了 Node 间容器网络的设置。在后续的 Pod 创建过程中, kubelet 将通过 CNI 接口调用 Calico 进行 Pod 网络的设置,包括 IP 地址、路由规则、Iptables 规则等。

如果设置 CALICO\_IPV4POOL\_IPIP="off",即不使用 IPIP 模式,则 Calico 将不会创建 tunl0 网络接口,路由规则直接使用物理机网卡作为路由器进行转发:

查看 k8s-node-1 服务器的路由表,可以看到一条到 k8s-node-2 的私网 10.1.140.64 的路由转发规则,将通过本机 ens33 网卡进行转发:

```
# ip route
default via 192.168.18.2 dev ens33
blackhole 10.1.109.64/26 proto bird
10.1.140.64/26 via 192.168.18.4 dev ens33 proto bird
172.17.1.0/24 dev docker0 proto kernel scope link src 172.17.1.1
192.168.18.0/24 dev ens33 proto kernel scope link src 192.168.18.3 metric 100
```

查看 k8s-node-2 服务器的路由表，可以看到一条到 k8s-node-1 的私网 10.1.109.64/26 的路由转发规则，将通过本机 ens33 网卡进行转发：

```
# ip route
default via 192.168.18.2 dev ens33
blackhole 10.1.140.64/26 proto bird
10.1.109.64/26 via 192.168.18.3 dev ens33 proto bird
172.17.2.0/24 dev docker0 proto kernel scope link src 172.17.2.1
192.168.18.0/24 dev ens33 proto kernel scope link src 192.168.18.4 metric 100
```

### 3) Calico 设置容器 IP 地址，跨主机容器网络连通性验证

下面我们创建几个 Pod，验证 Calico 对它们的网络设置。以第 1 章的 mysql 和 myweb 为例，分别创建 1 个 Pod 和两个 Pod：

```
mysql-rc.yaml
apiVersion: v1
kind: ReplicationController
metadata:
  name: mysql
spec:
  replicas: 1
  selector:
    app: mysql
  template:
    metadata:
      labels:
        app: mysql
    spec:
      containers:
        - name: mysql
          image: mysql
          ports:
            - containerPort: 3306
          env:
            - name: MYSQL_ROOT_PASSWORD
              value: "123456"
```

```
myweb-rc.yaml
apiVersion: v1
kind: ReplicationController
```



```

metadata:
  name: myweb
spec:
  replicas: 2
  selector:
    app: myweb
  template:
    metadata:
      labels:
        app: myweb
    spec:
      containers:
        - name: myweb
          image: kubeguide/tomcat-app:v1
          ports:
            - containerPort: 8080
          env:
            - name: MYSQL_SERVICE_HOST
              value: 'mysql'
            - name: MYSQL_SERVICE_PORT
              value: '3306'

```

```
# kubectl create -f mysql-rc.yaml -f myweb-rc.yaml
```

```
replicationcontroller "mysql" created
```

```
replicationcontroller "myweb" created
```

查看各 Pod 的 IP 地址，可以看到是通过 Calico 设置的以 10.1 开头的 IP 地址：

```
# kubectl get pod -o wide
```

NAME	READY	STATUS	RESTARTS	AGE	IP	NODE
mysql-8cztq	1/1	Running	0	2m	10.1.109.71	k8s-node-1
myweb-h4lg3	1/1	Running	0	2m	10.1.109.70	k8s-node-1
myweb-s86sk	1/1	Running	0	2m	10.1.140.66	k8s-node-2

进入运行在 k8s-node-2 的 Pod “myweb-s86sk”：

```
# kubectl exec -ti myweb-s86sk bash
```

在容器内访问运行在 k8s-node-1 上的 Pod “mysql-8cztq” 的 IP 地址 10.1.109.71：

```

root@myweb-s86sk:/usr/local/tomcat# ping 10.1.109.71
PING 10.1.109.71 (10.1.109.71): 56 data bytes
64 bytes from 10.1.109.71: icmp_seq=0 ttl=63 time=0.344 ms
64 bytes from 10.1.109.71: icmp_seq=1 ttl=63 time=0.213 ms

```

在容器内访问物理机 k8s-node-1 的 IP 地址 192.168.18.3：

```

root@myweb-s86sk:/usr/local/tomcat# ping 192.168.18.3
PING 192.168.18.3 (192.168.18.3): 56 data bytes
64 bytes from 192.168.18.3: icmp_seq=0 ttl=64 time=0.327 ms
64 bytes from 192.168.18.3: icmp_seq=1 ttl=64 time=0.182 ms

```

这说明跨主机容器间、容器与宿主机之间的网络都能互联互通了。

查看 k8s-node-2 物理机的网络接口和路由表，可以看到 Calico 为 Pod “myweb-s86sk” 新建了一个网络接口 cali439924adc43，并为其设置了一条路由规则：

```
# ip addr show
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN qlen 1
.....
7: cali439924adc43@if3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc noqueue
state UP
    link/ether e2:e9:9a:55:52:92 brd ff:ff:ff:ff:ff:ff link-netnsid 0
    inet6 fe80::e0e9:9aff:fe55:5292/64 scope link
        valid_lft forever preferred_lft forever

# ip route
default via 192.168.18.2 dev ens33
blackhole 10.1.140.64/26 proto bird
10.1.109.64/26 via 192.168.18.3 dev tunl0 proto bird onlink
10.1.140.66 dev cali439924adc43 scope link
172.17.2.0/24 dev docker0 proto kernel scope link src 172.17.2.1
192.168.18.0/24 dev ens33 proto kernel scope link src 192.168.18.4 metric 100
```

另外，Calico 还为该网络接口 cali439924adc43 设置了一系列 Iptables 规则：

```
# iptables -L
.....
Chain cali-from-wl-dispatch (2 references)
target prot opt source destination
cali-fw-cali439924adc43 all -- anywhere anywhere [goto]
/* cali:27N3bvAtjtNgABL_ */
DROP all -- anywhere anywhere /*
cali:tL986QdUS4OiW3mC */ /* Unknown interface */

Chain cali-fw-cali439924adc43 (1 references)
target prot opt source destination
ACCEPT all -- anywhere anywhere /*
cali:w_ft-rPVu6fgqGmc */ ctstate RELATED,ESTABLISHED
DROP all -- anywhere anywhere /*
cali:ATcF-FBghYxNthE2 */ ctstate INVALID
MARK all -- anywhere anywhere /*
cali:5mvqaVXl8wQh6vS6 */ MARK and 0xfeffffff
MARK all -- anywhere anywhere /*
cali:nOAdEHYzt1IeVaqu */ /* Start of policies */ MARK and 0xfdfdfdfdf

Chain cali-to-wl-dispatch (1 references)
target prot opt source destination
cali-tw-cali439924adc43 all -- anywhere anywhere [goto]
```

```
/* cali:WibRaHK-UmAeF88Y */

Chain cali-tw-cali439924adc43 (1 references)
  target      prot opt source                destination
ACCEPT      all  --  anywhere              anywhere             /*
cali:c2lcc_VY82hSFHuc */ ctstate RELATED,ESTABLISHED
DROP        all  --  anywhere              anywhere             /*
cali:6eNswYurPxc_1g2M */ ctstate INVALID
MARK        all  --  anywhere              anywhere             /*
cali:Y55YBsPr1TihN4NE */ MARK and 0xfeffffff
MARK        all  --  anywhere              anywhere             /*
cali:hfMD9kYf5exJluSH */ /* Start of policies */ MARK and 0xfdffffff
.....
```

#### 4) 使用网络策略实现 Pod 间的访问策略

下面以一个提供服务的 Nginx Pod 为例，为两个客户端 Pod 设置不同的网络访问权限，允许包含 Label “role=nginxclient” 的 Pod 访问 Nginx 容器，无此 Label 的其他容器则拒绝访问。为了实现这个需求，需要通过以下步骤来完成。

步骤 1: 首先为需要设置网络隔离的 Namespace 进行标注，本例中的所有 Pod 都在 Namespace default 中，故对其进行默认网络隔离的设置：

```
# kubectl annotate ns default
"net.beta.kubernetes.io/network-policy={\"ingress\": {\"isolation\":
\"DefaultDeny\"}}\"
```

设置完成后，default 内的各 Pod 之间的网络就无法连通了。

步骤 2: 创建 Nginx Pod，并添加 Label “app=nginx”：

```
nginx.yaml
apiVersion: v1
kind: Pod
metadata:
  name: nginx
  labels:
    app: nginx
spec:
  containers:
  - name: nginx
    image: nginx

# k create -f nginx.yaml
pod "nginx" created
```

步骤 3: 为 Nginx 设置准入访问策略，编辑文件 networkpolicy-allow-nginxclient.yaml，内容如下。

```
kind: NetworkPolicy
apiVersion: extensions/v1beta1
metadata:
  name: allow-nginxclient
spec:
  podSelector:
    matchLabels:
      app: nginx
  ingress:
    - from:
      - podSelector:
          matchLabels:
            role: nginxclient
        ports:
          - protocol: TCP
            port: 80
```

目标 Pod 应包含 Label “app=nginx”，允许访问的客户端 Pod 包含 Label “role=nginxclient”，并允许客户端访问 mysql 容器的 80 端口。

创建该 NetworkPolicy 资源对象：

```
# kubectl create -f networkpolicy-allow-nginxclient.yaml
networkpolicy "allow-nginxclient" created
```

步骤 4：创建两个客户端 Pod，一个包含 Label “role=nginxclient”，另一个无此 Label。分别进入各 Pod，访问 Nginx 容器，验证网络策略的效果。

```
client1.yaml
apiVersion: v1
kind: Pod
metadata:
  name: client1
  labels:
    role: nginxclient
spec:
  containers:
    - name: client1
      image: busybox
      command: [ "sleep", "3600" ]
```

```
client2.yaml
apiVersion: v1
kind: Pod
metadata:
  name: client2
spec:
  containers:
```

```
- name: client2
  image: busybox
  command: [ "sleep", "3600" ]

# kubectl create -f client1.yaml -f client2.yaml
pod "client1" created
pod "client2" created
```

登录 Pod “client1”:

```
# kubectl exec -ti client1 -- sh
```

尝试连接 Nginx 容器的 80 端口号:

```
/ # wget 10.1.109.69
Connecting to 10.1.109.69 (10.1.109.69:80)
index.html          100% |*****| 612  0:00:00 ETA
```

成功访问到 Nginx 的服务，说明 NetworkPolicy 生效。

登录 Pod “client2”:

```
# kubectl exec -ti client2 -- sh
```

尝试连接 Nginx 容器的 80 端口号:

```
/ # wget --timeout=5 10.1.109.69
Connecting to 10.1.109.69 (10.1.109.69:80)
wget: download timed out
```

访问超时，也说明 NetworkPolicy 生效，对没有 Label “role=nginxclient” 的客户端 Pod 拒绝访问。

本例中的网络策略是由 calico-policy-controller 具体实现的，calico-policy-controller 持续监听 Kubernetes 中 NetworkPolicy 的定义，与各 Pod 通过 Label 进行关联，将允许访问或拒绝访问的策略通知到各 calico-node 服务，最终 calico-node 完成对 Pod 间网络访问的设置，实现应用的网络隔离。

## 3.8 共享存储原理

### 3.8.1 共享存储机制概述

Kubernetes 对于有状态的容器应用或者对数据需要持久化的应用，不仅需要容器内的目录挂载到宿主机的目录或者 emptyDir 临时存储卷，而且需要更加可靠的存储来保存应用产生的重要数据，以便容器应用在重建之后，仍然可以使用之前的数据。不过，存储资源和计算资源

（CPU/内存）的管理方式完全不同。为了能够屏蔽底层存储实现的细节，让用户方便使用，同时能让管理员方便管理，Kubernetes 从 v1.0 版本就引入 PersistentVolume 和 PersistentVolumeClaim 两个资源对象来实现对存储的管理子系统。

PersistentVolume（PV）是对底层网络共享存储的抽象，将共享存储定义为一种“资源”，比如节点（Node）也是一种容器应用可以“消费”的资源。PV 由管理员进行创建和配置，它与共享存储的具体实现直接相关，例如 GlusterFS、iSCSI、RBD 或 GCE/AWS 公有云提供的共享存储，通过插件式的机制完成与共享存储的对接，以供应用访问和使用。

PersistentVolumeClaim（PVC）则是用户对于存储资源的一个“申请”。就像 Pod“消费”Node 的资源一样，PVC 会“消费”PV 资源。PVC 可以申请特定的存储空间和访问模式。

使用 PVC“申请”到一定的存储空间仍然不足以满足应用对于存储设备的各种需求。通常应用程序都会对存储设备的特性和性能有不同的要求，包括读写速度、并发性能、数据冗余等更高的要求，Kubernetes 从 v1.4 版本开始引入了一个新的资源对象 StorageClass，用于标记存储资源的特性和性能。到 v1.6 版本时，StorageClass 和动态资源供应的机制得到了完善，实现了存储卷的按需创建，在共享存储的自动化管理进程中实现了重要的一步。

通过 StorageClass 的定义，管理员可以将存储资源定义为某种类别（Class），正如存储设备对于自身的配置描述（Profile），例如“快速存储”“慢速存储”“有数据冗余”“无数据冗余”等。用户根据 StorageClass 的描述就能够直观得知各种存储资源的特性，就可以根据应用对存储资源的需求去申请存储资源了。

下面对 Kubernetes 的 PV、PVC、StorageClass 和动态资源供应等共享存储管理机制进行详细说明。

### 3.8.2 PV 详解

PV 作为存储资源，主要包括存储能力、访问模式、存储类型、回收策略、后端存储类型等关键信息的设置。下面的例子声明的 PV 具有如下属性：5Gi 存储空间，访问模式为“ReadWriteOnce”，存储类型为“slow”（要求系统中已存在名为 slow 的 StorageClass），回收策略为“Recycle”，并且后端存储类型为“nfs”（设置了 NFS Server 的 IP 地址和路径）：

```
apiVersion: v1
kind: PersistentVolume
metadata:
  name: pv1
spec:
  capacity:
    storage: 5Gi
  accessModes:
```

```
- ReadWriteOnce
persistentVolumeReclaimPolicy: Recycle
storageClassName: slow
nfs:
  path: /tmp
  server: 172.17.0.2
```

Kubernetes 支持的 PV 类型如下。

- ◎ `gcePersistentDisk`: GCE 公有云提供的 `PersistentDisk`。
- ◎ `AWSElasticBlockStore`: AWS 公有云提供的 `ElasticBlockStore`。
- ◎ `AzureFile`: Azure 公有云提供的 `File`。
- ◎ `AzureDisk`: Azure 公有云提供的 `Disk`。
- ◎ `FC` (Fibre Channel)。
- ◎ `Flocker`。
- ◎ `NFS`: 网络文件系统。
- ◎ `iSCSI`。
- ◎ `RBD` (Rados Block Device): Ceph 块存储。
- ◎ `CephFS`。
- ◎ `Cinder`: OpenStack Cinder 块存储。
- ◎ `GlusterFS`。
- ◎ `VsphereVolume`。
- ◎ `Quobyte Volumes`。
- ◎ `VMware Photon`。
- ◎ `Portworx Volumes`。
- ◎ `ScaleIO Volumes`。
- ◎ `HostPath`: 宿主机目录, 仅用于单机测试。

每种存储类型都有各自的特点, 在使用时需要根据它们各自的参数进行设置。

## 1. PV 的关键配置参数

### 1) 存储能力 (Capacity)

描述存储设备具备的能力, 目前仅支持对存储空间设置的 (storage=xx), 未来可能加入

IOPS、吞吐率等指标的设置。

2) 访问模式（Access Modes）

对 PV 进行访问模式的设置，用于描述用户应用对存储资源的访问的权限。访问模式如下。

- ◎ ReadWriteOnce（简写为 RWO）：读写权限，并且只能被单个 Node 挂载。
- ◎ ReadOnlyMany（简写为 ROX）：只读权限，允许被多个 Node 挂载。
- ◎ ReadWriteMany（简写为 RWX）：读写权限，允许被多个 Node 挂载。

某些 PV 可能支持多种访问模式，但 PV 在挂载时只能使用一种访问模式，多种访问模式不能同时生效。

表 3.6 描述了不同的存储提供者支持的访问模式，在 PV 的定义时需要与它们匹配。

表 3.6 不同的存储提供者支持的访问模式

Volume Plugin	ReadWriteOnce	ReadOnlyMany	ReadWriteMany
AWSElasticBlockStore	✓	-	-
AzureFile	✓	✓	✓
AzureDisk	✓	-	-
CephFS	✓	✓	✓
Cinder	✓	-	-
FC	✓	✓	-
FlexVolume	✓	✓	-
Flocker	✓	-	-
gcePersistentDisk	✓	✓	-
GlusterFS	✓	✓	✓
HostPath	✓	-	-
iSCSI	✓	✓	-
PhotonPersistentDisk	✓	-	-
Quobyte	✓	✓	✓
NFS	✓	✓	✓
RBD	✓	✓	-
VsphereVolume	✓	-	-
PortworxVolume	✓	-	✓
ScaleIO	✓	✓	-

3) 存储类别（Class）

PV 可以设定其存储的类别（Class），通过 storageClassName 参数指定一个 StorageClass 资



源对象的名称。具有特定“类别”的 PV 只能与请求了该“类别”的 PVC 进行绑定。未设定“类别”的 PV 则只能与不请求任何“类别”的 PVC 进行绑定。

#### 4) 回收策略 (Reclaim Policy)

目前支持如下三种回收策略。

- ◎ 保留 (Retain): 保留数据, 需要手工处理。
- ◎ 回收空间 (Recycle): 简单清除文件的操作 (例如执行 `rm -rf /thevolume/*` 命令)。
- ◎ 删除 (Delete): 与 PV 相连的后端存储完成 volume 的删除操作 (如 AWS EBS、GCE PD、Azure Disk、OpenStack Cinder 等设备的内部 volume 清理)。

目前, 只有 NFS 和 HostPath 两种类型的存储支持“Recycle”策略; AWS EBS、GCE PD、Azure Disk 和 Cinder volumes 支持“Delete”策略。

### 2. PV 生命周期的各个阶段 (Phase)

某个 PV 在生命周期中, 可能处于以下 4 个阶段之一。

- ◎ Available: 可用状态, 还未与某个 PVC 绑定。
- ◎ Bound: 已与某个 PVC 绑定。
- ◎ Released: 绑定的 PVC 已经删除, 资源已释放, 但没有被集群回收。
- ◎ Failed: 自动资源回收失败。

### 3. PV 的挂载参数 (Mount Options)

在将 PV 挂载到一个 Node 上时, 根据后端存储的特点, 可能需要设置额外的挂载参数, 目前可以通过在 PV 的定义中, 设置一个名为“`volume.beta.kubernetes.io/mount-options`”的 annotation 来实现。下面的例子对一个类型为 `gcePersistentDisk` 的 PV 设置了挂载参数“`discard`”:

```
apiVersion: "v1"
kind: "PersistentVolume"
metadata:
  name: gce-disk-1
  annotations:
    volume.beta.kubernetes.io/mount-options: "discard"
spec:
  capacity:
    storage: "10Gi"
  accessModes:
    - "ReadWriteOnce"
  gcePersistentDisk:
```

```
fsType: "ext4"  
pdName: "gce-disk-1"
```

并非所有类型的存储都支持设置挂载参数。从 Kubernetes v1.6 版本开始，以下存储类型支持设置挂载参数。

- ◎ gcePersistentDisk。
- ◎ AWSElasticBlockStore。
- ◎ AzureFile。
- ◎ AzureDisk。
- ◎ NFS。
- ◎ iSCSI。
- ◎ RBD（Rados Block Device）：Ceph 块存储。
- ◎ CephFS。
- ◎ Cinder：OpenStack 块存储。
- ◎ GlusterFS。
- ◎ VsphereVolume。
- ◎ Quobyte Volumes。
- ◎ VMware Photon。

定义了 PV 以后如何使用呢？这时就需要用到 PVC 了。下一节对 PVC 进行详细说明。

### 3.8.3 PVC 详解

PVC 作为用户对存储资源的需求申请，主要包括存储空间请求、访问模式、PV 选择条件和存储类别等信息的设置。下面的例子声明的 PVC 具有如下属性：申请 8Gi 存储空间，访问模式为“ReadWriteOnce”，PV 选择条件为包含标签“release=stable”并且包含条件为“environment In [dev]”的标签，存储类别为“slow”（要求系统中已存在名为 slow 的 StorageClass）：

```
kind: PersistentVolumeClaim  
apiVersion: v1  
metadata:  
  name: myclaim  
spec:  
  accessModes:  
    - ReadWriteOnce
```

```
resources:
  requests:
    storage: 8Gi
storageClassName: slow
selector:
  matchLabels:
    release: "stable"
  matchExpressions:
    - {key: environment, operator: In, values: [dev]}
```

PVC 的关键配置参数说明如下。

- ◎ **资源请求 (Resources):** 描述对存储资源的请求，目前仅支持 `request.storage` 的设置，即存储空间大小。
- ◎ **访问模式 (Access Modes):** PVC 也可以设置访问模式，用于描述用户应用对存储资源的访问权限。可以设置的三种访问模式与 PV 的设置相同。
- ◎ **PV 选择条件 (Selector):** 通过 Label Selector 的设置，可使 PVC 对于系统中已存在的各种 PV 进行筛选。系统将根据标签选择出合适的 PV 与该 PVC 进行绑定。选择条件可以使用 `matchLabels` 和 `matchExpressions` 进行设置，如果两个字段都设置了，则 Selector 的逻辑将是两组条件同时满足才能完成匹配。
- ◎ **存储类别 (Class):** PVC 在定义时可以设定需要的后端存储的“类别”（通过 `storageClassName` 字段指定），以降低对后端存储特性的详细信息的依赖。只有设置了该 Class 的 PV 才能被系统选出，并与该 PVC 进行绑定。

PVC 也可以不设置 Class 需求。如果 `storageClassName` 字段的值被设置为空 (`storageClassName=""`)，则表示该 PVC 不要求特定的 Class，系统将只选择未设定 Class 的 PV 与之匹配和绑定。PVC 也可以完全不设置 `storageClassName` 字段，此时将根据系统是否启用了名为“DefaultStorageClass”的 admission controller 进行相应的操作。

- ◎ 未启用 DefaultStorageClass：等效于 PVC 设置 `storageClassName` 的值为空 (`storageClassName=""`)，即只能选择未设定 Class 的 PV 与之匹配和绑定。
- ◎ 启用 DefaultStorageClass：要求集群管理员已定义默认的 StorageClass。如果系统中不存在默认的 StorageClass，则等效于不启用 DefaultStorageClass 的情况。如果存在默认的 StorageClass，则系统将自动为 PVC 创建一个 PV（使用默认 StorageClass 的后端存储），并将它们进行绑定。集群管理员设置默认 StorageClass 的方法为，在 StorageClass 的定义中加上一个 annotation “`storageclass.kubernetes.io/is-default-class=true`”。如果管理员将多个 StorageClass 都定义为 default，则由于不唯一，系统将无法为 PVC 创建相应的 PV。

注意，PVC 和 PV 都受限於 namespace，PVC 在选择 PV 时受到 namespace 的限制，只有相同 namespace 中的 PV 才可能与 PVC 绑定。Pod 在引用 PVC 时同样受 namespace 的限制，只有相同 namespace 中的 PVC 才能挂载到 Pod 内。

当 Selector 和 Class 都进行了设置时，系统将选择两个条件同时满足的 PV 与之匹配。

另外，如果资源供应使用的是动态模式，即管理员没有预先定义 PV，仅通过 StorageClass 交给系统自动完成 PV 的动态创建，那么 PVC 再设定 Selector 时，系统将无法为其供应任何存储资源了。

在启用动态供应模式的情况下，一旦用户删除了 PVC，与之绑定的 PV 将根据其默认的回收策略“Delete”也会被删除。如果需要保留 PV（用户数据），则在动态绑定成功后，用户需要将系统自动生成 PV 的回收策略从“Delete”改成“Retain”。

### 3.8.4 PV 和 PVC 的生命周期

PV 可以看作可用的存储资源，PVC 则是对存储资源的需求，PV 和 PVC 的相互关系遵循如图 3.45 所示的生命周期。

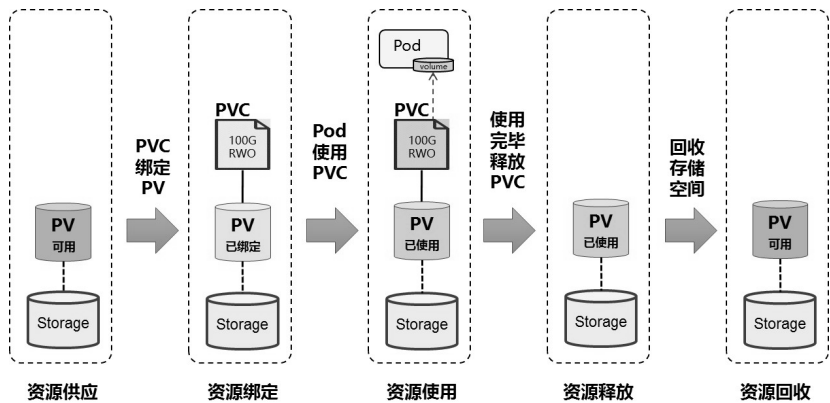


图 3.45 PV 和 PVC 的生命周期

#### 1. 资源供应（Provisioning）

Kubernetes 支持两种资源的供应模式：静态模式（Static）和动态模式（Dynamic）。资源供应的结果就是创建好的 PV。

- ◎ **静态模式：**集群管理员手工创建许多 PV，在定义 PV 时需要将后端存储的特性进行设置。

- ◎ **动态模式**：集群管理员无须手工创建 PV，而是通过 StorageClass 的设置对后端存储进行描述，标记为某种“类型（Class）”。此时要求 PVC 对存储的类型进行声明，系统将自动完成 PV 的创建及与 PVC 的绑定。PVC 可以声明 Class 为“”，说明该 PVC 禁止使用动态模式。

## 2. 资源绑定（Binding）

在用户定义好 PVC 之后，系统将根据 PVC 对存储资源的请求（存储空间和访问模式）在已存在的 PV 中选择一个满足 PVC 要求的 PV，一旦找到，就将该 PV 与用户定义的 PVC 进行绑定，然后用户的应用就可以使用这个 PVC 了。如果系统中没有满足 PVC 要求的 PV，PVC 则会无限期处于 Pending 状态，直到等到系统管理员创建了一个符合其要求的 PV。PV 一旦绑定到某个 PVC 上，就被这个 PVC 独占，不能再与其他 PVC 进行绑定了。在这种情况下，当 PVC 申请的存储空间比 PV 的少时，整个 PV 的空间都能够为 PVC 所用，可能会造成资源的浪费。如果资源供应使用的是动态模式，则系统在为 PVC 找到合适的 StorageClass 后，将自动创建一个 PV 并完成与 PVC 的绑定。

## 3. 资源使用（Using）

Pod 使用 volume 的定义，将 PVC 挂载到容器内的某个路径进行使用。volume 的类型为“persistentVolumeClaim”，在后面的示例中再进行详细说明。在容器应用挂载了一个 PVC 后，就能被持续独占使用。不过，多个 Pod 可以挂载同一个 PVC，应用程序需要考虑多个实例共同访问一块存储空间的问题。

## 4. 资源释放（Releasing）

当用户对存储资源使用完毕后，用户可以删除 PVC，与该 PVC 绑定的 PV 将会被标记为“已释放”，但还不能立刻与其他 PVC 进行绑定。通过之前 PVC 写入的数据可能还留在存储设备上，只有在清除之后该 PV 才能再次使用。

## 5. 资源回收（Reclaiming）

对于 PV，管理员可以设定回收策略（Reclaim Policy），用于设置与之绑定的 PVC 释放资源之后，对于遗留数据如何处理。只有 PV 的存储空间完成回收，才能供新的 PVC 绑定和使用。回收策略详见下节的说明。

下面通过两张图分别对在静态资源供应模式和动态资源供应模式下，PV、PVC、StorageClass 及 Pod 使用 PVC 的原理进行说明。

图 3.46 描述了在静态资源供应模式下，通过 PV 和 PVC 完成绑定，并供 Pod 使用的存储管理机制。

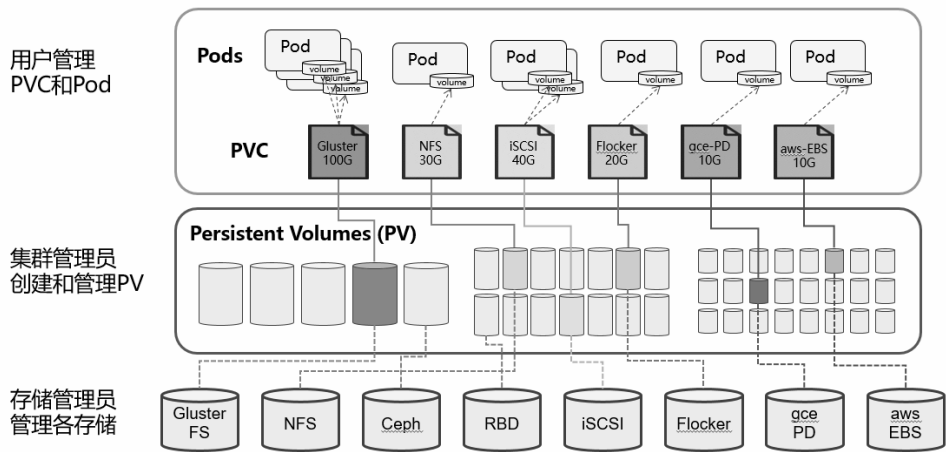


图 3.46 在静态模式下的 PV 和 PVC 原理

图 3.47 描述了在动态资源供应模式下，通过 StorageClass 和 PVC 完成资源动态绑定（系统自动生成 PV），并供 Pod 使用的存储管理机制。

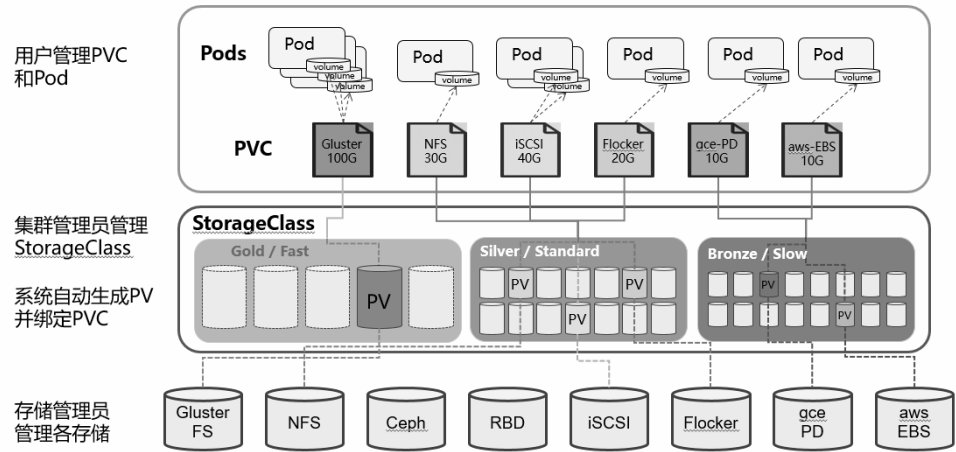


图 3.47 在动态模式下的 StorageClass、PV 和 PVC 原理

接下来，我们再看看 StorageClass 的概念和用法。

### 3.8.5 StorageClass 详解

StorageClass 作为对存储资源的抽象定义，对用户设置的 PVC 申请屏蔽后端存储的细节，一方面减轻用户对于存储资源细节的关注，另一方面也减轻了管理员手工管理 PV 的工作，由系统自动完成 PV 的创建和绑定，实现了动态的资源供应。使用基于 StorageClass 的动态资源供应模式将逐步成为云平台的标准存储配置模式。

StorageClass 的定义主要包括名称、后端存储的提供者（Provisioner）和后端存储的相关参数配置。StorageClass 一旦被创建出来，将无法修改。如需更改，则只能删除原 StorageClass 的定义重建。下面的例子定义了一个名为“standard”的 StorageClass，提供者 aws-ebs，其参数设置了一个 type=gp2。

```
kind: StorageClass
apiVersion: storage.k8s.io/v1
metadata:
  name: standard
provisioner: kubernetes.io/aws-ebs
parameters:
  type: gp2
```

#### 1. StorageClass 的关键配置参数

##### 1) 提供者（Provisioner）

描述存储资源的提供者，也可以看作后端存储驱动。目前 Kubernetes 支持的 Provisioner 都以“kubernetes.io/”为开头，用户也可以使用自定义的后端存储提供者。为了符合 StorageClass 的用法，自定义 Provisioner 需要符合存储卷的开发规范，详见 <https://github.com/kubernetes/community/blob/master/contributors/design-proposals/volume-provisioning.md> 的说明。

##### 2) 参数（Parameters）

后端存储资源提供者的参数设置，不同的 Provisioner 包括不同的参数设置。某些参数可以不显示设定，Provisioner 将使用其默认值。

接下来通过几种常见的 Provisioner 对 StorageClass 的定义进行详细说明。

##### ❖ AWS EBS 存储卷

```
kind: StorageClass
apiVersion: storage.k8s.io/v1
metadata:
  name: slow
provisioner: kubernetes.io/aws-ebs
parameters:
  type: io1
```

```
zone: us-east-1d
iopsPerGB: "10"
```

参数说明如下（详细说明请参考 AWS EBS 文档）。

- ◎ **type**: 可选项为 `io1`, `gp2`, `sc1`, `st1`, 默认值为 `gp2`。
- ◎ **zone**: AWS zone 的名称。
- ◎ **iopsPerGB**: 仅用于 `io1` 类型的 volume, 意为每秒每 GiB 的 I/O 操作数量。
- ◎ **encrypted**: 是否加密。
- ◎ **kmsKeyId**: 加密时的 Amazon Resource Name。

#### ❖ GCE PD 存储卷

```
kind: StorageClass
apiVersion: storage.k8s.io/v1
metadata:
  name: slow
provisioner: kubernetes.io/gce-pd
parameters:
  type: pd-standard
  zone: us-central1-a
```

参数说明如下（详细说明请参考 GCE 文档）。

- ◎ **type**: 可选项为 `pd-standard`, `pd-ssd`, 默认值为 `pd-standard`。
- ◎ **zone**: GCE zone 名称。

#### ❖ GlusterFS 存储卷

```
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: slow
provisioner: kubernetes.io/glusterfs
parameters:
  resturl: "http://127.0.0.1:8081"
  clusterid: "630372ccdc720a92c681fb928f27b53f"
  restauthenabled: "true"
  restuser: "admin"
  secretNamespace: "default"
  secretName: "heketi-secret"
  gidMin: "40000"
  gidMax: "50000"
  volumetype: "replicate:3"
```

参数说明如下（详细说明请参考 GlusterFS 和 Heketi 的文档）。



- ◎ `resturl`: Gluster REST 服务 (Heketi) 的 URL 地址, 用于自动完成 GlusterFSvolume 的设置。
- ◎ `restauthenabled`: 是否对 Gluster REST 服务启用安全机制。
- ◎ `restuser`: 访问 Gluster REST 服务的用户名。
- ◎ `secretNamespace` 和 `secretName`: 保存访问 Gluster REST 服务密码的 Secret 资源对象名。
- ◎ `clusterid`: GlusterFS 的 Cluster ID。
- ◎ `gidMin` 和 `gidMax`: StorageClass 的 GID 范围, 用于动态资源供应时为 PV 设置的 GID。
- ◎ `volumetype`: GlusterFS 的 volume 类型设置, 例如 `replicate:3` (Replicate 类型, 3 份副本); `disperse:4:2` (Disperse 类型, 数据 4 份, 冗余 2 份; “none” (Distribute 类型)。

#### ❖ OpenStack Cinder 存储卷

```
kind: StorageClass
apiVersion: storage.k8s.io/v1
metadata:
  name: gold
provisioner: kubernetes.io/cinder
parameters:
  type: fast
  availability: nova
```

参数说明如下。

- ◎ `type`: Cinder 的 VolumeType, 默认值为空。
- ◎ `availability`: Availability Zone, 默认值为空。

其他 Provisioner 的 StorageClass 相关参数设置请参考它们各自的配置手册。

## 2. 设置默认的 (Default) StorageClass

要在系统中设置一个默认的 StorageClass, 首先需要启用名为 “DefaultStorageClass” 的 admission controller, 即在 kube-apiserver 的命令行参数 `--admission-control` 中增加:

```
--admission-control=...,DefaultStorageClass
```

然后, 在 StorageClass 的定义中设置一个 annotation:

```
kind: StorageClass
apiVersion: storage.k8s.io/v1
metadata:
  name: gold
  annotations:
    storageclass.beta.kubernetes.io/is-default-class="true"
```

```
provisioner: kubernetes.io/gce-pd
parameters:
  type: pd-ssd
```

通过 `kubectl create` 命令创建成功后，查看 `StorageClass` 列表，可以看到名为 `gold` 的 `StorageClass` 被标记为 “default”：

```
# kubectl get sc
NAME          TYPE
gold (default) kubernetes.io/gce-pd
```

### 3.8.6 动态存储管理实战：GlusterFS

本节以 `GlusterFS` 为例，从定义 `StorageClass`、创建 `GlusterFS` 和 `Heketi` 服务、用户申请 `PVC` 到创建 `Pod` 使用存储资源，对 `StorageClass` 和动态资源分配进行详细说明，进一步剖析 `Kubernetes` 的存储机制。

#### 1. 准备工作

为了能够使用 `GlusterFS`，首先在计划用于 `GlusterFS` 的各 `Node` 上安装 `GlusterFS` 客户端：

```
# yum install glusterfs glusterfs-fuse
```

`GlusterFS` 管理服务容器需要以特权模式运行，在 `kube-apiserver` 的启动参数中增加：

```
--allow-privileged=true
```

给要部署 `GlusterFS` 管理服务的节点打上“`storagenode=glusterfs`”的标签，是为了将 `GlusterFS` 容器定向部署到安装了 `GlusterFS` 的 `Node`：

```
# kubectl label node k8s-node-1 storagenode=glusterfs
# kubectl label node k8s-node-2 storagenode=glusterfs
# kubectl label node k8s-node-3 storagenode=glusterfs
```

#### 2. 创建 `GlusterFS` 管理服务容器集群

`GlusterFS` 管理服务容器以 `Daemonset` 的方式进行部署，确保每台 `Node` 上都运行一个 `GlusterFS` 管理服务。`glusterfs-daemonset.yaml` 内容如下：

```
kind: DaemonSet
apiVersion: extensions/v1beta1
metadata:
  name: glusterfs
  labels:
    glusterfs: daemonset
  annotations:
```

```

description: GlusterFS DaemonSet
tags: glusterfs
spec:
  template:
    metadata:
      name: glusterfs
      labels:
        glusterfs-node: pod
    spec:
      nodeSelector:
        storagenode: glusterfs
      hostNetwork: true
      containers:
        - image: gluster/gluster-centos:latest
          name: glusterfs
          volumeMounts:
            - name: glusterfs-heketi
              mountPath: "/var/lib/heketi"
            - name: glusterfs-run
              mountPath: "/run"
            - name: glusterfs-lvm
              mountPath: "/run/lvm"
            - name: glusterfs-etc
              mountPath: "/etc/glusterfs"
            - name: glusterfs-logs
              mountPath: "/var/log/glusterfs"
            - name: glusterfs-config
              mountPath: "/var/lib/glusterd"
            - name: glusterfs-dev
              mountPath: "/dev"
            - name: glusterfs-misc
              mountPath: "/var/lib/misc/glusterfsd"
            - name: glusterfs-cgroup
              mountPath: "/sys/fs/cgroup"
              readOnly: true
            - name: glusterfs-ssl
              mountPath: "/etc/ssl"
              readOnly: true
          securityContext:
            capabilities: {}
            privileged: true
      readinessProbe:
        timeoutSeconds: 3
        initialDelaySeconds: 60
        exec:
          command:

```

```
- "/bin/bash"
- "-c"
- systemctl status glusterd.service
livenessProbe:
  timeoutSeconds: 3
  initialDelaySeconds: 60
  exec:
    command:
      - "/bin/bash"
      - "-c"
      - systemctl status glusterd.service
volumes:
- name: glusterfs-heketi
  hostPath:
    path: "/var/lib/heketi"
- name: glusterfs-run
- name: glusterfs-lvm
  hostPath:
    path: "/run/lvm"
- name: glusterfs-etc
  hostPath:
    path: "/etc/glusterfs"
- name: glusterfs-logs
  hostPath:
    path: "/var/log/glusterfs"
- name: glusterfs-config
  hostPath:
    path: "/var/lib/glusterd"
- name: glusterfs-dev
  hostPath:
    path: "/dev"
- name: glusterfs-misc
  hostPath:
    path: "/var/lib/misc/glusterfsd"
- name: glusterfs-cgroup
  hostPath:
    path: "/sys/fs/cgroup"
- name: glusterfs-ssl
  hostPath:
    path: "/etc/ssl"
# kubectl create -f glusterfs-daemonset.yaml
daemonset "glusterfs" created
# kubectl get po
NAME                READY    STATUS    RESTARTS   AGE
glusterfs-k2src     1/1      Running   0           1m
glusterfs-q32z2     1/1      Running   0           1m
```

### 3. 创建 Heketi 服务

Heketi 是一个提供 RESTful API 管理 GlusterFS 卷的框架,并能够在 OpenStack、Kubernetes、OpenShift 等云平台上实现动态存储资源供应,支持 GlusterFS 多集群管理,便于管理员对 GlusterFS 进行操作。图 3.48 简单描述了 Heketi 的作用。

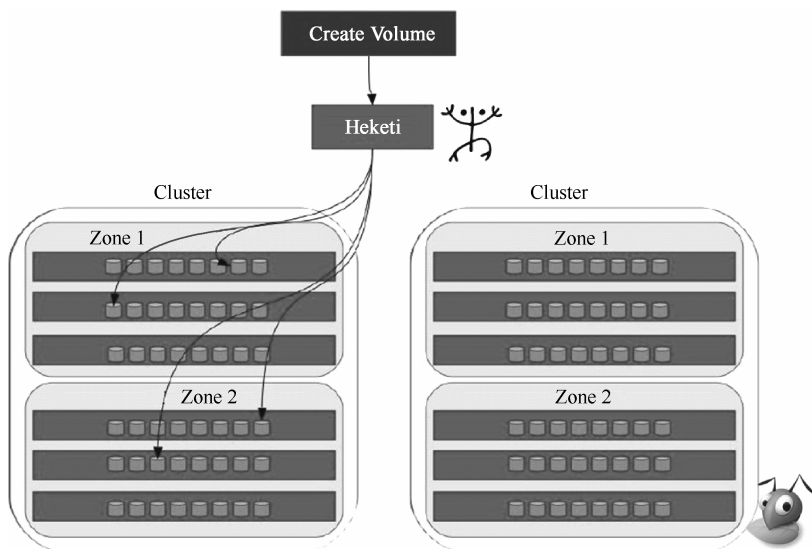


图 3.48 Heketi 的作用

在部署 Heketi 服务之前,需要为它创建一个 ServiceAccount 对象:

**heketi-service-account.yaml**

```
apiVersion: v1
kind: ServiceAccount
metadata:
  name: heketi-service-account
```

```
# kubectl create -f heketi-service-account.yaml
serviceaccount "heketi-service-account" created
```

部署 Heketi 服务:

**heketi-deployment-svc.yaml**

```
---
kind: Deployment
apiVersion: extensions/v1beta1
metadata:
  name: deploy-heketi
  labels:
    glusterfs: heketi-deployment
```

```
    deploy-heketi: heketi-deployment
  annotations:
    description: Defines how to deploy Heketi
spec:
  replicas: 1
  template:
    metadata:
      name: deploy-heketi
      labels:
        name: deploy-heketi
        glusterfs: heketi-pod
    spec:
      serviceAccountName: heketi-service-account
      containers:
        - image: heketi/heketi:dev
          name: deploy-heketi
          env:
            - name: HEKETI_EXECUTOR
              value: kubernetes
            - name: HEKETI_FSTAB
              value: "/var/lib/heketi/fstab"
            - name: HEKETI_SNAPSHOT_LIMIT
              value: '14'
            - name: HEKETI_KUBE_GLUSTER_DAEMONSET
              value: "y"
          ports:
            - containerPort: 8080
          volumeMounts:
            - name: db
              mountPath: "/var/lib/heketi"
          readinessProbe:
            timeoutSeconds: 3
            initialDelaySeconds: 3
            httpGet:
              path: "/hello"
              port: 8080
          livenessProbe:
            timeoutSeconds: 3
            initialDelaySeconds: 30
            httpGet:
              path: "/hello"
              port: 8080
      volumes:
        - name: db
          hostPath:
            path: "/heketi-data"
---
```

```

kind: Service
apiVersion: v1
metadata:
  name: deploy-heketi
  labels:
    glusterfs: heketi-service
    deploy-heketi: support
  annotations:
    description: Exposes Heketi Service
spec:
  selector:
    name: deploy-heketi
  ports:
    - name: deploy-heketi
      port: 8080
      targetPort: 8080

```

需要注意的是，Heketi 的 db 数据需要持久化保存，建议使用 hostPath 或其他共享存储进行保存。

```

# kubectl create -f heketi-deployment-svc.yaml
deployment "deploy-heketi" created
service "deploy-heketi" created

```

#### 4. 为 Heketi 设置 GlusterFS 集群

在 Heketi 能够管理 GlusterFS 集群之前，首先要为其设置 GlusterFS 集群的信息。可以用一个 topology.json 配置文件来完成各个 GlusterFS 节点和设备的定义。Heketi 要求一个 GlusterFS 集群中至少有 3 个节点。在 topology.json 配置文件的 hostnames 字段中的 manage 上填写主机名，在 storage 上填写 IP 地址，devices 要求为未创建文件系统的裸设备（可以有多块盘），以供 Heketi 自动完成 PV (Physical volume)、VG (Volume group) 和 LV (Logical volume) 的创建。topology.json 文件的内容如下：

```

{
  "clusters": [
    {
      "nodes": [
        {
          "node": {
            "hostnames": {
              "manage": [
                "k8s-node-1"
              ],
              "storage": [
                "192.168.18.3"
              ]
            }
          }
        }
      ]
    }
  ]
}

```

```
    },
    "zone": 1
  },
  "devices": [
    "/dev/sdb"
  ]
},
{
  "node": {
    "hostnames": {
      "manage": [
        "k8s-node-2"
      ],
      "storage": [
        "192.168.18.4"
      ]
    },
    "zone": 1
  },
  "devices": [
    "/dev/sdb"
  ]
},
{
  "node": {
    "hostnames": {
      "manage": [
        "k8s-node-3"
      ],
      "storage": [
        "192.168.18.5"
      ]
    },
    "zone": 1
  },
  "devices": [
    "/dev/sdb"
  ]
}
]
}
```

进入 Heketi 容器，使用命令行工具 `heketi-cli` 完成 GlusterFS 集群的创建：

```
# export HEKETI_CLI_SERVER=http://localhost:8080
# heketi-cli topology load --json=topology.json
```



```

Creating cluster ... ID: f643da1cd64691c5705932a46a95d1d5
  Creating node k8s-node-1 ... ID: 883506b091a22bd13f10bc3d0fb51223
    Adding device /dev/sdb ... OK
  Creating node k8s-node-2 ... ID: e64b879689106f82a9c4ac910a865cc8
    Adding device /dev/sdb ... OK
  Creating node k8s-node-3 ... ID: b7783484180f6a592a30baebfb97d9be
    Adding device /dev/sdb ... OK

```

经过这个操作，Heketi 完成了 GlusterFS 集群的创建，同时在 GlusterFS 集群的各个节点的 /dev/sdb 盘上成功创建了 PV（Physical volume）和 VG（Volume group）。

查看 Heketi 的 topology 信息，可以看到 Node 和 Device 的详细信息，包括磁盘空间的大小和剩余空间。此时，Volumes 和 Bricks 还未创建：

```

# heketi-cli topology info
Cluster Id: f643da1cd64691c5705932a46a95d1d5

Volumes:

Nodes:

Node Id: 883506b091a22bd13f10bc3d0fb51223
State: online
Cluster Id: f643da1cd64691c5705932a46a95d1d5
Zone: 1
Management Hostname: k8s-node-1
Storage Hostname: 192.168.18.3
Devices:
  Id:b474f14b0903ed03ec80d4a989f943f2  Name:/dev/sdb
State:online  Size (GiB):9  Used (GiB):0  Free (GiB):9
  Bricks:

Node Id: b7783484180f6a592a30baebfb97d9be
State: online
Cluster Id: f643da1cd64691c5705932a46a95d1d5
Zone: 1
Management Hostname: k8s-node-3
Storage Hostname: 192.168.18.5
Devices:
  Id:fac3fa5ac1de3d5bde3aa68f6aa61285  Name:/dev/sdb
State:online  Size (GiB):9  Used (GiB):0  Free (GiB):9
  Bricks:

Node Id: e64b879689106f82a9c4ac910a865cc8
State: online
Cluster Id: f643da1cd64691c5705932a46a95d1d5
Zone: 1

```

```
Management Hostname: k8s-node-2
Storage Hostname: 192.168.18.4
Devices:
    Id:05532e7db723953e8643b64b36aee1d1    Name:/dev/sdb
State:online    Size (GiB):9    Used (GiB):0    Free (GiB):9
Bricks:
```

## 5. 定义 StorageClass

准备工作已经就绪，集群管理员现在可以在 Kubernetes 集群中定义一个 StorageClass 了。  
storageclass-gluster-heketi.yaml 配置文件的内容如下：

```
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: gluster-heketi
provisioner: kubernetes.io/glusterfs
parameters:
  resturl: "http://172.17.2.2:8080"
  restauthenabled: "false"
```

Provisioner 参数须设置为 “kubernetes.io/glusterfs”。

resturl 的地址需要设置为 API Server 所在主机可以访问到的 Heketi 服务的某个地址，可以使用服务 ClusterIP+端口号、容器 IP 地址+端口号或将服务映射到物理机，使用物理机 IP+NodePort。

创建这个 StorageClass 资源对象：

```
# kubectl create -f storageclass-gluster-heketi.yaml
storageclass "gluster-heketi" created
```

## 6. 定义 PVC

现在，用户可以申请一个 PVC 了。例如一个用户申请一个 1Gi 空间的共享存储资源，StorageClass 使用 “gluster-heketi”，未定义任何 Selector，说明使用动态资源供应模式。

```
pvc-gluster-heketi.yaml
kind: PersistentVolumeClaim
apiVersion: v1
metadata:
  name: pvc-gluster-heketi
spec:
  storageClassName: gluster-heketi
  accessModes:
    - ReadWriteOnce
  resources:
```

```

requests:
  storage: 1Gi
# k create -f pvc-gluster-heketi.yaml
persistentvolumeclaim "pvc-gluster-heketi" created

```

PVC 的定义一旦生成，系统便将触发 Heketi 进行相应的操作，主要为在 GlusterFS 集群上创建 brick，再创建并启动一个 volume。整个过程可以在 Heketi 的日志中查到：

```

.....
[kubeexec] DEBUG 2017/04/26 00:51:30
/src/github.com/heketi/heketi/executors/kubeexec/kubeexec.go:250: Host:
k8s-node-1 Pod: glusterfs-ld7nh Command: gluster --mode=script volume create
vol_87b9314cb76bafacfb7e9cdc04fc05 replica 3
192.168.18.3:/var/lib/heketi/mounts/vg_b474f14b0903ed03ec80d4a989f943f2/brick_d0
8520c9ff7b9a0a9165f9815671f2cd/brick
192.168.18.5:/var/lib/heketi/mounts/vg_fac3fa5ac1de3d5bde3aa68f6aa61285/brick_68
18dce118b8a54e9590199d44a3817b/brick
192.168.18.4:/var/lib/heketi/mounts/vg_05532e7db723953e8643b64b36ae1d1/brick_9e
cb8f7fdela937011f04401e7c6c56/brick
Result: volume create: vol_87b9314cb76bafacfb7e9cdc04fc05: success: please
start the volume to access data
.....
[kubeexec] DEBUG 2017/04/26 00:51:33
/src/github.com/heketi/heketi/executors/kubeexec/kubeexec.go:250: Host:
k8s-node-1 Pod: glusterfs-ld7nh Command: gluster --mode=script volume start
vol_87b9314cb76bafacfb7e9cdc04fc05
Result: volume start: vol_87b9314cb76bafacfb7e9cdc04fc05: success
.....

```

查看 PVC 的状态，可见其已经为 Bound（已绑定）：

```

# kubectl get pvc
NAME                                STATUS    VOLUME                                CAPACITY
ACCESSMODES  STORAGECLASS  AGE
pvc-gluster-heketi  Bound        pvc-783cf949-2a1a-11e7-8717-000c29eaed40  1Gi
RWX                gluster-heketi  6m

```

查看 PV，可见系统自动创建的 PV：

```

# kubectl get pv
NAME                                CAPACITY  ACCESSMODES  RECLAIMPOLICY
STATUS  CLAIM                                STORAGECLASS  REASON  AGE
pvc-783cf949-2a1a-11e7-8717-000c29eaed40  1Gi        RWX          Delete
Bound   default/pvc-gluster-heketi  gluster-heketi                6m

```

查看该 PV 的详细信息，可以看到其容量、引用的 StorageClass 等信息都已正确设置，状态也为 Bound（已绑定），回收策略则为默认的“Delete”。同时 Gluster 的 Endpoint 和 Path 也由 Heketi 自动完成了设置。

```

# kubectl describe pv pvc-783cf949-2a1a-11e7-8717-000c29eaed40

```

```
Name:          pvc-783cf949-2a1a-11e7-8717-000c29eaed40
Labels:        <none>
Annotations:   pv.beta.kubernetes.io/gid=2000
               pv.kubernetes.io/bound-by-controller=yes
               pv.kubernetes.io/provisioned-by=kubernetes.io/glusterfs
StorageClass:  gluster-heketi
Status:        Bound
Claim:         default/pvc-gluster-heketi
Reclaim Policy: Delete
Access Modes:  RWX
Capacity:      1Gi
Message:
Source:
  Type:        Glusterfs (a Glusterfs mount on the host that shares a pod's
lifetime)
EndpointsName: glusterfs-dynamic-pvc-gluster-heketi
Path:          vol_87b9314cb76bafacfb7e9cdc04fcac05
ReadOnly:      false
Events:        <none>
```

至此，一个可供 Pod 使用的 PVC 就创建成功了。接下来 Pod 就能通过 volume 的设置将这个 PVC 挂载到容器内部进行使用了。

## 7. Pod 使用 PVC 的存储资源

在 Pod 中使用 PVC 定义的存储资源将非常容易，只需设置一个 volume，其类型为 persistentVolumeClaim，即可轻松引用一个 PVC。下例中使用一个 busybox 容器验证对 PVC 的使用，注意 Pod 需要与 PVC 属于同一个 namespace：

```
pod-use-pvc.yaml
apiVersion: v1
kind: Pod
metadata:
  name: pod-use-pvc
spec:
  containers:
  - name: pod-use-pvc
    image: busybox
    command:
    - sleep
    - "3600"
    volumeMounts:
    - name: gluster-volume
      mountPath: "/pv-data"
      readOnly: false
  volumes:
```

```
- name: gluster-volume
  persistentVolumeClaim:
    claimName: pvc-gluster-heketi

# kubectl create -f pod-use-pvc.yaml
pod "pod-use-pvc" created
```

进入容器 `pod-use-pvc`，在 `/pv-data` 目录下创建一些文件：

```
# kubectl exec -ti pod-use-pvc -- /bin/sh
/ # cd /pv-data
/ # touch a
/ # echo "hello" > b
```

可以验证文件 `a` 和 `b` 在 `GlusterFS` 集群中正确生成。

至此，使用 `Kubernetes` 最新的动态存储供应模式，配合 `StorageClass` 和 `Heketi` 共同搭建基于 `GlusterFS` 的共享存储就完成了。有兴趣的读者可以继续尝试 `StorageClass` 的其他设置，例如调整 `GlusterFS` 的 `volume` 类型、修改 `PV` 的回收策略等。

在使用动态存储供应模式的情况下，相对于静态模式的优势至少包括如下两点。

(1) 管理员无须预先创建大量的 `PV` 作为存储资源。

(2) 用户申请 `PVC` 时无法保证容量与预置 `PV` 的容量完全匹配。从 `Kubernetes v1.6` 开始，建议用户优先考虑使用 `StorageClass` 的动态存储供应模式进行存储管理。

# 第 4 章

## Kubernetes 开发指南

---

本章将引入 REST 的概念, 详细说明 Kubernetes API, 并举例说明如何基于 Jersey 和 Fabric8 框架访问 Kubernetes API, 深入分析基于这两个框架访问 Kubernetes API 的优缺点。下面从 REST 开始说起。

### 4.1 REST 简述

---

REST (Representational State Transfer) 是由 Roy Thomas Fielding 博士在他的论文 *Architectural Styles and the Design of Network-based Software Architectures* 中提出的一个术语。REST 本身只是为分布式超媒体系统设计的一种架构风格, 而不是标准。

基于 Web 的架构实际上就是各种规范的集合, 这些规范共同组成了 Web 架构, 比如 HTTP、客户端服务器模式都是规范。每当我们在原有规范的基础上增加新的规范时, 就会形成新的架构。而 REST 正是这样一种架构, 它结合了一系列规范, 形成了一种新的基于 Web 的架构风格。

传统的 Web 应用大多是 B/S 架构, 涉及如下规范。

(1) 客户-服务器: 这种规范的提出, 改善了用户接口跨多个平台的可移植性, 并且通过简化服务器组件, 改善了系统的可伸缩性。最为关键的是通过分离用户接口和数据存储, 使得不同的用户终端共享相同的数据成为可能。

(2) 无状态性: 无状态性是在客户-服务器约束的基础上添加的又一层规范, 它要求通信必须在本质上是无状态的, 即从客户端到服务器的每个 request 都必须包含理解该 request 所必需

的所有信息。这个规范改善了系统的可见性（无状态性使得客户端和服务端不必保存对方的详细信息，服务器只需要处理当前的 `request`，而不必了解所有 `request` 的历史）、可靠性（无状态性减少了服务器从局部错误中恢复的任务量）、可伸缩性（无状态性使得服务器端可以很容易地释放资源，因为服务器端不必在多个 `request` 中保存状态）。同时，这种规范的缺点也是显而易见的，由于不能将状态数据保存在服务器上，因此增加了在一系列 `request` 中发送重复数据的开销，严重降低了效率。

（3）缓存：为了改善无状态性带来的网络的低效性，我们添加了缓存约束。缓存约束允许隐式或显式地标记一个 `response` 中的数据，赋予了客户端缓存 `response` 数据的功能，这样就可以为以后的 `request` 共用缓存的数据，部分或全部地消除一部分交互，提高了网络效率。但是由于客户端缓存了信息，所以增加了客户端与服务器数据不一致的可能性，从而降低了可靠性。

B/S 架构的优点是部署非常方便，在用户体验方面却不很理想。为了改善这种情况，我们引入了 REST。REST 在原有架构上增加了三个新规范：统一接口、分层系统和按需代码。

（1）统一接口：REST 架构风格的核心特征就是强调组件之间有一个统一的接口，表现为在 REST 世界里，网络上的所有事物都被抽象为资源，REST 通过通用的链接器接口对资源进行操作。这样设计的好处是保证系统提供的服务都是解耦的，极大地简化了系统，从而改善了系统的交互性和可重用性。

（2）分层系统：分层系统规则的加入提高了各种层次之间的独立性，为整个系统的复杂性设置了边界，通过封装遗留的服务，使新的服务器免受遗留客户端的影响，也提高了系统的可伸缩性。

（3）按需代码：REST 允许对客户端功能进行扩展。比如，通过下载并执行 `applet` 或脚本形式的代码来扩展客户端的功能。但这在改善系统可扩展性的同时降低了可见性，所以它只是 REST 的一个可选约束。

REST 架构是针对 Web 应用而设计的，其目的是为了降低开发的复杂性，提高系统的可伸缩性。REST 提出了如下设计准则。

- （1）网络上的所有事物都被抽象为资源（Resource）。
- （2）每个资源对应一个唯一的资源标识符（Resource Identifier）。
- （3）通过通用的连接器接口（Generic Connector Interface）对资源进行操作。
- （4）对资源的各种操作不会改变资源标识符。
- （5）所有的操作都是无状态的（Stateless）。

REST 中的资源所指的并不是数据，而是数据和表现形式的组合，比如“最新访问的 10 位会

员”和“最活跃的 10 位会员”在数据上可能有重叠或者完全相同，而由于它们的表现形式不同，所以被归为不同的资源，这也就是为什么 REST 的全名是 Representational State Transfer。资源标识符就是 URI（Uniform Resource Identifier），不管是图片、Word 还是视频文件，甚至只是一种虚拟的服务，也不管是 xml、txt 还是其他文件格式，全部通过 URI 对资源进行唯一标识。

REST 是基于 HTTP 的，任何对资源的操作行为都通过 HTTP 来实现。以往的 Web 开发大多数用的是 HTTP 中的 GET 和 POST 方法，很少使用其他方法，这实际上是因为对 HTTP 的片面理解造成的。HTTP 不仅仅是一个简单的运载数据的协议，而且是一个具有丰富内涵的网络软件的协议，它不仅能对互联网资源进行唯一定位，还能告诉我们如何对该资源进行操作。HTTP 把对一个资源的操作限制在 4 种方法内：GET、POST、PUT 和 DELETE，这正是对资源 CRUD 操作的实现。由于资源和 URI 是一一对应的，在执行这些操作时 URI 没有变化，和以往的 Web 开发有很大的区别，所以极大地简化了 Web 开发，也使得 URI 可以被设计成更为直观地反映资源的结构。这种 URI 的设计被称作 RESTful 的 URI，为开发人员引入了一种新的思维方式：通过 URL 来设计系统结构。当然了，这种设计方式对于一些特定情况也是不适用的，也就是说不是所有 URI 都适用于 RESTful。

REST 之所以可以提高系统的可伸缩性，就是因为它要求所有操作都是无状态的。由于没有了上下文（Context）的约束，做分布式和集群时就更为简单，也可以让系统更为有效地利用缓冲池（Pool），并且由于服务器端不需要记录客户端的一系列访问，也就减少了服务器端的性能损耗。

Kubernetes API 也符合 RESTful 规范，下面对其进行介绍。

## 4.2 Kubernetes API 详解

### 4.2.1 Kubernetes API 概述

Kubernetes API 是集群系统中的重要组成部分，Kubernetes 中各种资源（对象）的数据通过该 API 接口被提交到后端的持久化存储（etcd）中，Kubernetes 集群中的各部件之间通过该 API 接口实现解耦合，同时 Kubernetes 集群中一个重要且便捷的管理工具 kubectl 也是通过访问该 API 接口实现其强大的管理功能的。Kubernetes API 中的资源对象都拥有通用的元数据，资源对象也可能存在嵌套现象，比如在一个 Pod 里面嵌套多个 Container。创建一个 API 对象是指通过 API 调用创建一条有意义的记录，该记录一旦被创建，Kubernetes 将确保对应的资源对象会被自动创建并托管维护。



在 Kubernetes 系统中,大多数情况下,API 定义和实现都符合标准的 HTTP REST 格式,比如通过标准的 HTTP 动词 (POST、PUT、GET、DELETE) 来完成对相关资源对象的查询、创建、修改、删除等操作。但同时 Kubernetes 也为某些非标准的 REST 行为实现了附加的 API 接口,例如 Watch 某个资源的变化、进入容器执行某个操作等。另外,某些 API 接口可能违背严格的 REST 模式,因为接口不是返回单一的 JSON 对象,而是返回其他类型的数据,比如 JSON 对象流 (Stream) 或非结构化的文本日志数据等。

Kubernetes 开发人员认为,任何成功的系统都会经历一个不断成长和不断适应各种变更的过程。因此,他们期望 Kubernetes API 是不断变更和增长的。同时,他们在设计和开发时,有意识地兼容了已存在的客户需求。通常,新的 API 资源 (Resource) 和新的资源域不希望被频繁地加入系统。资源或域的删除需要一个严格的审核流程。

Kubernetes API 文档官网为 <https://kubernetes.io/docs/api-reference/v1.6>。

运行在 Master 节点上的 API Server 进程同时提供了 Swagger 格式的 API 网页。Swagger UI 是一款 REST API 文档在线自动生成和功能测试软件,关于 Swagger 的内容请访问官网 <http://swagger.io>。

从 Kubernetes v1.6 开始,需要设置 kube-apiserver 的启动参数 `--enable-swagger-ui=true` 启用这个页面,其访问地址为 `http://<master-ip>:<master-port>/swagger-ui/`。假设 API Server 启动在 192.168.18.3 服务器上的 8080 端口,则可以通过访问 `http://192.168.1.128:8080/swagger-ui/` 来查看 API 列表,如图 4.1 所示。

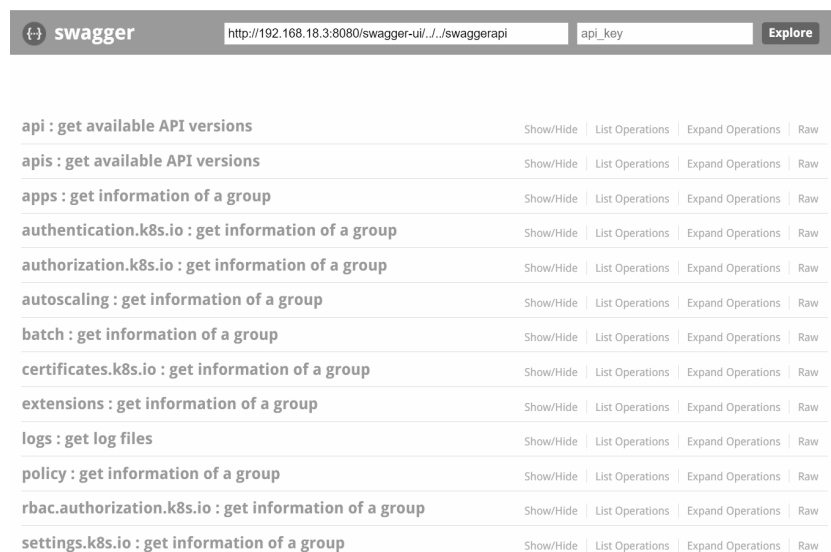


图 4.1 API Server 自带的 swagger-ui 页面

单击 `api/v1` 可以查看所有 API 的列表，如图 4.2 所示。

v1: API at /api/v1		Show/Hide	List Operations	Expand Operations	Raw
GET	/api/v1	get available resources			
GET	/api/v1/componentstatuses	list objects of kind ComponentStatus			
GET	/api/v1/componentstatuses/{name}	read the specified ComponentStatus			
GET	/api/v1/configmaps	list or watch objects of kind ConfigMap			
GET	/api/v1/endpoints	list or watch objects of kind Endpoints			
GET	/api/v1/events	list or watch objects of kind Event			
GET	/api/v1/limitranges	list or watch objects of kind LimitRange			
DELETE	/api/v1/namespaces	delete collection of Namespace			
GET	/api/v1/namespaces	list or watch objects of kind Namespace			
POST	/api/v1/namespaces	create a Namespace			
POST	/api/v1/namespaces/{namespace}/bindings	create a Binding			
DELETE	/api/v1/namespaces/{namespace}/configmaps	delete collection of ConfigMap			
GET	/api/v1/namespaces/{namespace}/configmaps	list or watch objects of kind ConfigMap			
POST	/api/v1/namespaces/{namespace}/configmaps	create a ConfigMap			
DELETE	/api/v1/namespaces/{namespace}/configmaps/{name}	delete a ConfigMap			

图 4.2 查看 API 列表

以创建一个 Pod 为例，找到 Rest API 的访问路径为：`/api/v1/namespaces/{namespace}/pods`，如图 4.3 所示。

POST	/api/v1/namespaces/{namespace}/pods	create a Pod
------	-------------------------------------	--------------

图 4.3 创建一个 Pod

单击链接展开，即可查看详细的 API 接口说明，如图 4.4 所示。

POST	/api/v1/namespaces/{namespace}/pods	create a Pod
Response Class (Status )		
Model   Model Schema		
<pre>{   "kind": "",   "apiVersion": "",   "metadata": {     "name": "",     "generateName": "",     "namespace": "",     "selfLink": "",     "uid": ""   } }</pre>		
Response Content Type: application/json		
Parameters		
Parameter	Value	Description   Parameter Type   Data Type
pretty	(empty)	If 'true', then the output is pretty printed.   query   string

图 4.4 详细的 API 接口说明

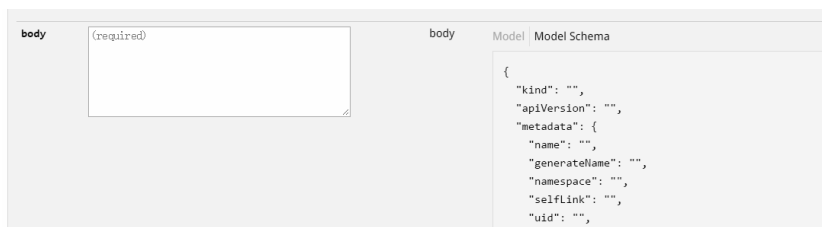


图 4.4 详细的 API 接口说明（附）

单击 **Model** 链接，则可以查看文本格式显示的 API 接口描述，如图 4.5 所示。

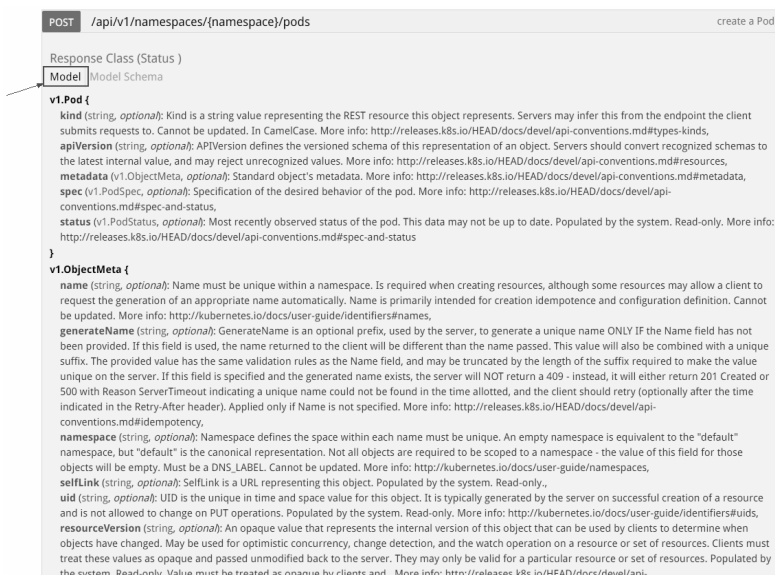


图 4.5 API 接口描述

我们看到，在 Kubernetes API 中，一个 API 的顶层（Top Level）元素由 kind、apiVersion、metadata、spec 和 status 等几个部分组成，接下来，我们分别对这几个部分进行说明。

kind 表明对象有以下三大类别。

（1）对象（objects）：代表在系统中的一个永久资源（实体），例如 Pod、RC、Service、Namespace 及 Node 等。通过操作这些资源的属性，客户端可以对该对象进行创建、修改、删除和获取操作。

（2）列表（list）：一个或多个资源类别的集合。列表有一个通用元数据的有限集合。所有列表（lists）通过“items”域获得对象数组，例如 PodLists、ServiceLists、NodeLists。大部分定义在系统中的对象都有一个返回所有资源（resource）集合的端点，以及零到多个返回所有资源

集合的子集的端点。某些对象有可能是单例对象（singletons），例如当前用户、系统默认用户等，这些对象没有列表。

（3）简单类别（simple）：该类别包含作用在对象上的特殊行为和非持久实体。该类别限制了使用范围，它有一个通用元数据的有限集合，例如 **Binding**、**Status**。

**apiVersion** 表明 API 的版本号，当前版本默认只支持 **v1**。

**Metadata** 是资源对象的元数据定义，是集合类的元素类型，包含一组由不同名称定义的属性。在 **Kubernetes** 中每个资源对象都必须包含以下 3 种 **Metadata**。

（1）**namespace**：对象所属的命名空间，如果不指定，系统则会将对象置于名为“**default**”的系统命名空间中。

（2）**name**：对象的名字，在一个命名空间中名字应具备唯一性。

（3）**uid**：系统为每个对象生成的唯一 ID，符合 **RFC 4122** 规范的定义。

此外，每种对象还应该包含以下几个重要元数据。

（1）**labels**：用户可定义的“标签”，键和值都为字符串的 **map**，是对对象进行组织和分类的一种手段，通常用于标签选择器（**Label Selector**），用来匹配目标对象。

（2）**annotations**：用户可定义的“注解”，键和值都为字符串的 **map**，被 **Kubernetes** 内部进程或者某些外部工具使用，用于存储和获取关于该对象的特定元数据。

（3）**resourceVersion**：用于识别该资源内部版本号的字符串，在用于 **Watch** 操作时，可以避免在 **GET** 操作和下一次 **Watch** 操作之间造成的信息不一致，客户端可以用它来判断资源是否改变。该值应该被客户端看作不透明，且不做任何修改就返回给服务端。客户端不应该假定版本信息具有跨命名空间、跨不同资源类别、跨不同服务器的含义。

（4）**creationTimestamp**：系统记录创建对象时的时间戳，符合 **RFC 3339** 规范。

（5）**deletionTimestamp**：系统记录删除对象时的时间戳，符合 **RFC 3339** 规范。

（6）**selfLink**：通过 API 访问资源自身的 URL，例如一个 Pod 的 link 可能是 `/api/v1/namespaces/default/pods/frontend-o8bg4`。

**spec** 是集合类的元素类型，用户对需要管理的对象进行详细描述的主体部分都在 **spec** 里给出，它会被 **Kubernetes** 持久化到 **etcd** 中保存，系统通过 **spec** 的描述来创建或更新对象，以达到用户期望的对象运行状态。**spec** 的内容既包括用户提供的配置设置、默认值、属性的初始化值，也包括在对象创建过程中由其他相关组件（例如 **schedulers**、**auto-scalers**）创建或修改的对象属性，比如 Pod 的 **Service IP** 地址。如果 **spec** 被删除，那么该对象将会从系统中被删除。

**Status** 用于记录对象在系统中的当前状态信息，它也是集合类元素类型，**status** 在一个自动

处理的进程中被持久化，可以在流转的过程中生成。如果观察到一个资源丢失了它的状态 (Status)，则该丢失的状态可能被重新构造。以 Pod 为例，Pod 的 status 信息主要包括 conditions、containerStatuses、hostIP、phase、podIP、startTime 等。其中比较重要的两个状态属性如下。

(1) phase: 描述对象所处的生命周期阶段，phase 的典型值是 Pending (创建中)、Running、Active (正在运行中) 或 Terminated (已终结)，这几种状态对于不同的对象可能有轻微的差别，此外，关于当前 phase 附加的详细说明可能包含在其他域中。

(2) condition: 表示条件，由条件类型和状态值组成，目前仅有一种条件类型 Ready，对应的状态值可以为 True、False 或 Unknown。一个对象可以具备多种 condition，而 condition 的状态值也可能不断发生变化，condition 可能附带一些信息，例如最后的探测时间或最后的转变时间。

### 4.2.2 API 版本

为了在兼容旧版本的同时不断升级新的 API，Kubernetes 提供了多版本 API 的支持能力，每个版本的 API 通过一个版本号路径前缀进行区分，例如/api/v1beta3。通常情况下，新旧几个不同的 API 版本都能涵盖所有的 Kubernetes 资源对象，在不同的版本之间这些 API 接口存在一些细微差别。Kubernetes 开发团队基于 API 级别选择版本而不是基于资源和域级别，是为了确保 API 能够描述一个清晰的连续的系统资源和行为的视图，能够控制访问的整个过程和控制实验性 API 的访问。

API 及版本发布建议描述了版本升级的当前思路。版本 v1beta1、v1beta2 和 v1beta3 为不建议使用 (Deprecated) 的版本，请尽快转到 v1 版本。在 2015 年 6 月 4 日，Kubernetes v1 版本 API 正式发布。版本 v1beta1 和 v1beta2 API 在 2015 年 6 月 1 日被删除，版本 v1beta3 API 在 2015 年 7 月 6 日被删除。

### 4.2.3 API Groups (API 组)

为了更容易对 API 进行扩展，Kubernetes 使用 API Groups (API 组) 进行标识。API Groups 以 REST URL 中的路径进行定义。当前支持两类 API groups。

- ◎ Core Groups (核心组)，也可以称为 Legacy Groups，作为 Kubernetes 最核心的 API，其特点是没有“组”的概念，例如“v1”，在资源对象的定义中表示为“apiVersion: v1”。
- ◎ 具有分组信息的 API，以/apis/\$GROUP\_NAME/\$VERSION URL 路径进行标识，在资源对象的定义中表示为“apiVersion: \$GROUP\_NAME/\$VERSION”，例如：“apiVersion: batch/v1”“apiVersion: extensions:v1beta1”“apiVersion: apps/v1beta1”等，详细的 API

列表参见官网 <https://kubernetes.io/docs/reference>，目前根据 Kubernetes 的不同版本有不同的 API 说明页面。

例如 Pod 的 API 说明如图 4.6 所示，由于 Pod 属于核心资源对象，所以不存在某个扩展 API Group，页面显示为“Core”，在 Pod 的定义中为“apiVersion: v1”。

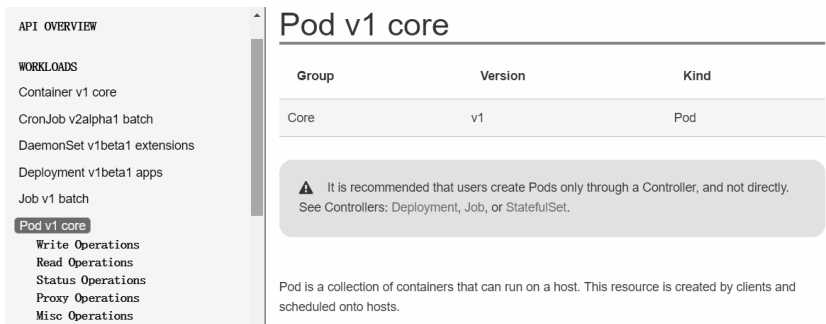


图 4.6 Pod 的 API 说明

而 StatefulSet 则属于名为“apps”的 API 组，版本号为 v1beta1，在 StatefulSet 的定义中为“apiVersion: apps/v1beta1”，如图 4.7 所示。

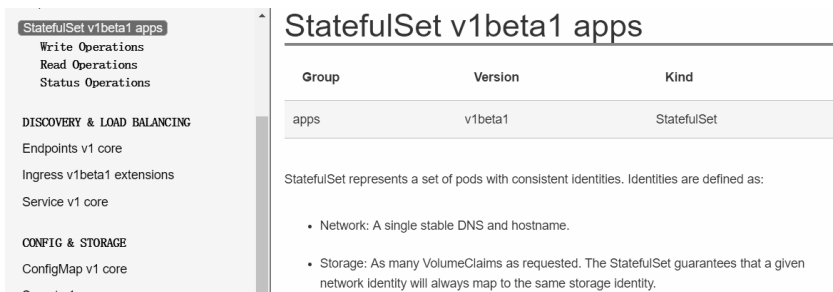


图 4.7 StatefulSet 的 API 说明

如果要启用或禁用特定的 API 组，则需要在 API Server 的启动参数中设置--runtime-config 进行声明，例如--runtime-config=batch/v2alpha1 表示启用 API 组“batch/v2alpha1”，也可以设置--runtime-config=batch/v1=false 表示禁用 API 组“batch/v1”。多个 API 组的设置以逗号分隔。在当前的 API Server 服务中，DaemonSets、Deployments、HorizontalPodAutoscalers、Ingress、Jobs 和 ReplicaSets 所属的 API 组是默认启用的。

在未来的演进中，Kubernetes 将支持用户自定义资源对象和对它们的基本 CRUD 操作。伴随着自定义资源对象，Kubernetes 的未来版本还将支持用户自定义 API Server，与 Kubernetes Master 的 API Server 共同工作，完成更复杂的资源对象管理工作。

### 4.2.4 API 方法说明

API 资源使用 REST 模式，对资源对象的操作方法如下。

(1) GET /<资源名的复数格式>: 获得某一类型的资源列表，例如 GET /pods 返回一个 Pod 资源列表。

(2) POST /<资源名的复数格式>: 创建一个资源，该资源来自用户提供的 JSON 对象。

(3) GET /<资源名复数格式>/<名字>: 通过给出的名称 (Name) 获得单个资源，例如 GET /pods/first 返回一个名称为 “first” 的 Pod。

(4) DELETE /<资源名复数格式>/<名字>: 通过给出的名字删除单个资源，在删除选项 (DeleteOptions) 中可以指定优雅删除 (Grace Deletion) 的时间 (GracePeriodSeconds)，该可选项表明了从服务端接收到删除请求到资源被删除的时间间隔 (单位为 s)。不同的类别 (Kind) 可能为优雅删除时间 (Grace Period) 申明默认值。用户提交的优雅删除时间将覆盖该默认值，包括值为 0 的优雅删除时间。

(5) PUT /<资源名复数格式>/<名字>: 通过给出的资源名和客户端提供的 JSON 对象来更新或创建资源。

(6) PATCH /<资源名复数格式>/<名字>: 选择修改资源详细指定的域。

对于 PATCH 操作，目前 Kubernetes API 通过相应的 HTTP 首部 “Content-Type” 对其进行识别。

目前支持以下三种类型的 PATCH 操作。

(1) JSON Patch, Content-Type: application/json-patch+json。在 RFC6902 的定义中，JSON Patch 是执行在资源对象上的一系列操作，例如 {"op": "add", "path": "/a/b/c", "value": ["foo", "bar"]}。详情请查看 RFC6902 说明，网址为 <https://tools.ietf.org/html/rfc6902>。

(2) Merge Patch, Content-Type: application/merge-json-patch+json。在 RFC7386 的定义中，Merge Patch 必须包含对一个资源对象的部分描述，这个资源对象的部分描述就是一个 JSON 对象。该 JSON 对象被提交到服务端，并和服务端的当前对象合并，从而创建一个新的对象。详情请查看 RFC7386 说明，网址为 <https://tools.ietf.org/html/rfc7386>。

(3) Strategic Merge Patch, Content-Type: application/strategic-merge-patch+json。Strategic Merge Patch 是一个定制化的 Merge Patch 实现。接下来将详细讲解 Strategic Merge Patch。

在标准的 JSON Merge Patch 中，JSON 对象总是被合并 (merge) 的，但是资源对象中的列表域总是被替换的。通常这不是用户所希望的。例如，我们通过下列定义创建一个 Pod 资源对象：

```
spec:
  containers:
    - name: nginx
      image: nginx-1.0
```

接着我们希望添加一个容器到这个 Pod 中，代码和上传的 JSON 对象如下所示：

```
PATCH /api/v1/namespaces/default/pods/pod-name
spec:
  containers:
    - name: log-tailer
      image: log-tailer-1.0
```

如果我们使用标准的 Merge Patch，则其中的整个容器列表将被单个的“log-tailer”容器所替换。然而我们的目的是两个容器列表能够合并。

为了解决这个问题，Strategic Merge Patch 通过添加元数据到 API 对象中，并通过这些新元数据来决定哪个列表被合并，哪个列表不被合并。当前这些元数据作为结构标签，对于 API 对象自身来说是合法的。对于客户端来说，这些元数据作为 Swagger annotations 也是合法的。在上述例子中，向“containers”中添加“patchStrategy”域，且它的值为“merge”，通过添加“patchMergeKey”，它的值为“name”。也就是说，“containers”中的列表将会被合并而不是替换，合并的依据为“name”域的值。

此外，Kubernetes API 添加了资源变动的“观察者”模式的 API 接口。

- ◎ GET /watch/<资源名复数格式>: 随时间变化，不断接收一连串的 JSON 对象，这些 JSON 对象记录了给定资源类别内所有资源对象的变化情况。
- ◎ GET /watch/<资源名复数格式>/<name>: 随时间变化，不断接收一连串的 JSON 对象，这些 JSON 对象记录了某个给定资源对象的变化情况。

上述接口改变了返回数据的基本类别，watch 动词返回的是一连串的 JSON 对象，而不是单个的 JSON 对象。并不是所有的对象类别都支持“观察者”模式的 API 接口，在后续的章节中将会说明哪些资源对象支持这种接口。

另外，Kubernetes 还增加了 HTTP Redirect 与 HTTP Proxy 这两种特殊的 API 接口，前者实现资源重定向访问，后者则实现 HTTP 请求的代理。

#### 4.2.5 API 响应说明

API Server 响应用户请求时附带一个状态码，该状态码符合 HTTP 规范。表 4.1 列出了 API Server 可能返回的状态码。



表 4.1 API Server 可能返回的状态码

状 态 码	编 码	描 述
200	OK	表明请求完全成功
201	Created	表明创建类的请求完全成功
204	NoContent	表明请求完全成功，同时 HTTP 响应不包含响应体。 在响应 OPTIONS 方法的 HTTP 请求时返回
307	TemporaryRedirect	表明请求资源的地址被改变，建议客户端使用 Location 首部给出的临时 URL 来定位资源
400	BadRequest	表明请求是非法的，建议客户不要重试，修改该请求
401	Unauthorized	表明请求能够到达服务端，且服务端能够理解用户请求，但是拒绝做更多的事情，因为客户端必须提供认证信息。如果客户端提供了认证信息，则返回该状态码，表明服务端指出所提供的认证信息不合适或非法
403	Forbidden	表明请求能够到达服务端，且服务端能够理解用户请求，但是拒绝做更多的事情，因为该请求被设置成拒绝访问。建议客户不要重试，修改该请求
404	NotFound	表明所请求的资源不存在。建议客户不要重试，修改该请求
405	MethodNotAllowed	表明请求中带有该资源不支持的方法。建议客户不要重试，修改该请求
409	Conflict	表明客户端尝试创建的资源已经存在，或者由于冲突请求的更新操作不能被完成
422	UnprocessableEntity	表明由于所提供的作为请求部分的数据非法，创建或修改操作不能被完成
429	TooManyRequests	表明超出了客户端访问频率的限制或者服务端接收到多于它能处理的请求。建议客户端读取相应的 Retry-After 首部，然后等待该首部指出的时间后再重试
500	InternalServerError	表明服务端能被请求访问到，但是不能理解用户的请求；或者服务端内产生非预期中的一个错误，而且该错误无法被认知；或者服务端不能在一个合理的时间内完成处理（这可能由于服务器临时负载过重造成或者由于和其他服务器通信时的一个临时通信故障造成）
503	ServiceUnavailable	表明被请求的服务无效。建议客户不要重试修改该请求
504	ServerTimeout	表明请求在给定的时间内无法完成。客户端仅在为请求指定超时（Timeout）参数时会得到该响应

在调用 API 接口发生错误时，Kubernetes 将会返回一个状态类别（Status Kind）。下面是两种常见的错误场景。

（1）当一个操作不成功时（例如，当服务端返回一个非 2xx HTTP 状态码时）。

（2）当一个 HTTP DELETE 方法调用失败时。

状态对象被编码成 JSON 格式，同时该 JSON 对象被作为请求的响应体。该状态对象包含人和机器使用的域，这些域中包含来自 API 的关于失败原因的详细信息。状态对象中的信息补充了对 HTTP 状态码的说明。例如：

```
$ curl -v -k -H "Authorization: Bearer WhCDvq4VPpYhrccmF6ei7V9qlbqTubUc"
HTTps://10.240.122.184:443/api/v1/namespaces/default/pods/grafana
> GET /api/v1/namespaces/default/pods/grafana HTTP/1.1
```

```
> User-Agent: curl/7.26.0
> Host: 10.240.122.184
> Accept: */*
> Authorization: Bearer WhCDvq4VPpYhrCFmF6ei7V9qlbqTubUc
>

< HTTP/1.1 404 Not Found
< Content-Type: application/json
< Date: Wed, 20 May 2015 18:10:42 GMT
< Content-Length: 232
<
{
  "kind": "Status",
  "apiVersion": "v1",
  "metadata": {},
  "status": "Failure",
  "message": "pods \"grafana\" not found",
  "reason": "NotFound",
  "details": {
    "name": "grafana",
    "kind": "pods"
  },
  "code": 404
}
```

- ◎ “status” 域包含两个可能的值：Success 和 Failure。
- ◎ “message” 域包含对错误的可读描述。
- ◎ “reason” 域包含说明该操作失败原因的可读描述。如果该域的值为空，则表示该域内没有任何说明信息。“reason” 域澄清 HTTP 状态码，但没有覆盖该状态码。
- ◎ “details” 可能包含和 “reason” 域相关的扩展数据。每个 “reason” 域可以定义它的扩展的 “details” 域。该域是可选的，返回数据的格式是不确定的，不同的 reason 类型返回的 “details” 域的内容不一样。

## 4.3 使用 Java 程序访问 Kubernetes API

本节介绍如何使用 Java 程序访问 Kubernetes API。在 Kubernetes 的官网上列出了多个访问 Kubernetes API 的开源项目，其中有两个是用 Java 语言开发工具的开源项目，一个是 OSGL，另一个是 Fabric8。在本节所列的两个 Java 开发例子中，一个是基于 Jersey 的，另一个是基于 Fabric8 的。

### 4.3.1 Jersey

Jersey 是一个 RESTful 请求服务 JAVA 框架。与 Struts 类似，它可以和 Hibernate、Spring 框架整合。通过它不仅方便开发 RESTful Web Service，而且可以将它作为客户端方便地访问 RESTful Web Service 服务端。

如果没有一个好的工具包，则开发一个能够用不同的媒介（Media）类型无缝地暴露你的数据，以及很好地抽象客户、服务端通信的底层通信的 RESTful Web Services，会很不容易。为了能够简化用 Java 开发 RESTful Web Service 及其客户端的流程，业界设计了 JAX-RS API。Jersey RESTful Web Services 框架是一个开源的高质量框架，它为用 JAVA 语言开发 RESTful Web Service 及其客户端而生，支持 JAX-RS APIs。Jersey 不仅支持 JAX-RS APIs，而且在此基础上扩展了 API 接口，这些扩展更加方便和简化了 RESTful Web Services 及其客户端的开发。

由于 Kubernetes API Server 是 RESTful Web Service，因此此处选用 Jersey 框架开发 RESTful Web Service 客户端，用来访问 Kubernetes API。在本例中选用的 Jersey 框架的版本为 1.19，所涉及的 Jar 包如图 4.8 所示。









 commons-codec-1.2.jar	2015/9/13 11:10	Executable Jar File	30 KB
 commons-httpclient-3.1.jar	2015/9/13 11:09	Executable Jar File	298 KB
 commons-logging-1.0.4.jar	2015/9/13 11:10	Executable Jar File	38 KB
 jackson-core-asl-1.9.2.jar	2015/2/11 5:41	Executable Jar File	223 KB
 jackson-jaxrs-1.9.2.jar	2015/2/11 5:41	Executable Jar File	18 KB
 jackson-mapper-asl-1.9.2.jar	2015/2/11 5:41	Executable Jar File	748 KB
 jackson-xc-1.9.2.jar	2015/2/11 5:41	Executable Jar File	27 KB
 jersey-apache-client-1.19.jar	2015/2/11 5:41	Executable Jar File	22 KB
 jersey-atom-abdera-1.19.jar	2015/2/11 5:41	Executable Jar File	20 KB
 jersey-client-1.19.jar	2015/2/11 5:41	Executable Jar File	131 KB
 jersey-core-1.19.jar	2015/2/11 5:41	Executable Jar File	427 KB
 jersey-guice-1.19.jar	2015/2/11 5:41	Executable Jar File	16 KB
 jersey-json-1.19.jar	2015/2/11 5:41	Executable Jar File	162 KB
 jersey-multipart-1.19.jar	2015/2/11 5:41	Executable Jar File	53 KB
 jersey-server-1.19.jar	2015/2/11 5:41	Executable Jar File	687 KB
 jersey-servlet-1.19.jar	2015/2/11 5:41	Executable Jar File	126 KB
 jersey-simple-server-1.19.jar	2015/2/11 5:41	Executable Jar File	12 KB
 jersey-spring-1.19.jar	2015/2/11 5:41	Executable Jar File	18 KB
 jettison-1.1.jar	2015/2/11 5:41	Executable Jar File	67 KB
 jsr311-api-1.1.1.jar	2015/2/11 5:41	Executable Jar File	46 KB
 oauth-client-1.19.jar	2015/2/11 5:41	Executable Jar File	15 KB
 oauth-server-1.19.jar	2015/2/11 5:41	Executable Jar File	30 KB
 oauth-signature-1.19.jar	2015/2/11 5:41	Executable Jar File	24 KB

图 4.8 本例所涉及的 Jar 包

对 Kubernetes API 的访问包含如下三个方面。

(1) 指明访问资源的类型。

(2) 访问时的一些选项（参数），比如命名空间、对象的名称、过滤方式（标签和域）、子目录、访问的目标是否是代理和是否用 `watch` 方式访问等。

(3) 访问的方法，比如增、删、改、查。

在使用 Jersey 框架访问 Kubernetes API 之前，为这三个方面定义了三个对象。第 1 个定义的对象为 `ResourceType`，它定义了访问资源的类型；第 2 个定义的对象是 `Params`，它定义了访问 API 时的一些选项，以及通过这些选项如何生成完整的 URI；第 3 个定义的对象是 `RestfulClient`，它是一个接口，该接口定义了访问 API 的方法（Method）。

`ResourceType` 是一个 `ENUM` 类型的对象，定义了 16 种资源，代码如下：

```
package com.hp.k8s.apiclient.imp;

public enum ResourceType {
    NODES("nodes"),
    NAMESPACES("namespaces"),
    SERVICES("services"),
    REPLICATIONCONTROLLERS("replicationcontrollers"),
    PODS("pods"),
    BINDINGS("bindings"),
    ENDPOINTS("endpoints"),
    SERVICEACCOUNTS("serviceaccounts"),
    SECRETS("secrets"),
    EVENTS("events"),
    COMPONENTSTATUSES("componentstatuses"),
    LIMITRANGES("limitranges"),
    RESOURCEQUOTAS("resourcequotas"),
    PODTEMPLATES("podtemplates"),
    PERSISTENTVOLUMECLAIMS("persistentvolumeclaims"); PERSISTENTVOLUMES
("persistentvolumes");
    private String type;

    private ResourceType(String type) {
        this.type = type;
    }

    public String getType() {
        return type;
    }
}
```

`Params` 对象的代码如下：

```
package com.hp.k8s.apiclient.imp;
```

```

import java.io.UnsupportedEncodingException;
import java.net.URLEncoder;
import java.util.List;
import java.util.Map;

import org.apache.logging.log4j.LogManager;
import org.apache.logging.log4j.Logger;

public class Params {
    private static final Logger LOG = LogManager.getLogger(Params.class.getName());
    private String namespace = null;
    private String name = null;
    private Map<String, String> fields = null;
    private Map<String, String> labels = null;
    private Map<String, String> notLabels = null;
    private Map<String, List<String>> inLabels = null;
    private Map<String, List<String>> notInLabels = null;
    private String json = null;
    private ResourceType resourceType = null;
    private String subPath = null;
    private boolean isVisitProxy = false;
    private boolean isSetWatcher = false;

    public String buildPath() {
        StringBuilder result = (isVisitProxy ? new StringBuilder("/proxy")
            : (isSetWatcher ? new StringBuilder("/watch") : new
StringBuilder(""));
        if (null != namespace)
            result.append("/namespaces/").append(namespace);

        result.append("/").append(resourceType.getType());
        if (null != name)
            result.append("/").append(name);
        if (null != subPath)
            result.append("/").append(subPath);

        if (null != labels && !labels.isEmpty() || null != notLabels && !notLabels.
isEmpty()
            || null != inLabels && inLabels.size() > 0 || null != notInLabels
&& notInLabels.size() > 0
            || null != fields && fields.size() > 0) {
            StringBuilder labelSelectorStr = null;
            StringBuilder fieldSelectorStr = null;
            try {
                labelSelectorStr = builderLabelSelector();
                fieldSelectorStr = builderFiledSelector();
            } catch (UnsupportedEncodingException e1) {

```

```
        LOG.error(e1);
    }

    if (labelSelectorStr.length() + fieldSelectorStr.length() > 0)
        result.append("?");
    if (labelSelectorStr.length() > 0) {
        result.append("labelSelector="). append(labelSelectorStr.
toString());

        if (fieldSelectorStr.length() > 0) {
            result.append(",");
        }
    }
    if (fieldSelectorStr.length() > 0) {
        result.append("fieldSelector=").append(fieldSelectorStr.
toString());
    }

}

return result.toString();
}

private StringBuilder builderLabelSelector() throws UnsupportedEncoding
Exception {
    StringBuilder result = new StringBuilder();
    if (null != labels) {
        for (String key : labels.keySet()) {
            if (result.length() > 0) {
                result.append(",");
            }

            result.append(URLEncoder.encode(key + "=" + labels.get(key),
"GBK"));
        }
    }

    if (null != notLabels) {
        for (String key : labels.keySet()) {
            if (result.length() > 0) {
                result.append(",");
            }

            result.append(URLEncoder.encode(key + "!=" + labels.get(key),
"GBK"));
        }
    }
}
```

```

        if (null != inLabels) {
            for (String key : inLabels.keySet()) {
                if (result.length() > 0) {
                    result.append(URLEncoder.encode(",", "GBK"));
                }
                result.append(URLEncoder.encode(key + " in (" + listToString
(inLabels.get(key), ",") + ")", "GBK"));
            }
        }

        if (null != notInLabels) {
            for (String key : inLabels.keySet()) {
                if (result.length() > 0) {
                    result.append(URLEncoder.encode(",", "GBK"));
                }
                result.append(URLEncoder.encode(key + " notin (" + listToString
(inLabels.get(key), ",") + ")", "GBK"));
            }
        }

        LOG.info("label result: " + result);
        return result;
    }

    private StringBuilder builderFiledSelector() throws UnsupportedEncodingException
Exception {
        StringBuilder result = new StringBuilder();
        if (null != fields) {
            for (String key : fields.keySet()) {
                if (result.length() > 0) {
                    result.append(",");
                }

                result.append(URLEncoder.encode(key + "=" + fields.get(key),
"GBK"));
            }
        }

        return result;
    }

    private String listToString(List<String> list, String delim) {
        boolean isFirst = true;
        StringBuilder result = new StringBuilder();
        for (String str : list) {
            if (isFirst) {

```

```
        result.append(str);
        isFirst = false;
    } else {
        result.append(delim).append(str);
    }
}

return result.toString();
}

public String getNamespace() {
    return namespace;
}

public void setNamespace(String namespace) {
    this.namespace = namespace;
}

public String getName() {
    return name;
}

public void setName(String name) {
    this.name = name;
}

public Map<String, String> getFields() {
    return fields;
}

public void setFields(Map<String, String> fields) {
    this.fields = fields;
}

public Map<String, String> getLabels() {
    return labels;
}

public void setLabels(Map<String, String> labels) {
    this.labels = labels;
}

public String getJson() {
    return json;
}

public void setJson(String json) {
```



```

        this.json = json;
    }

    public ResourceType getResourceType() {
        return resourceType;
    }

    public void setResourceType(ResourceType resourceType) {
        this.resourceType = resourceType;
    }

    public String getSubPath() {
        return subPath;
    }

    public void setSubPath(String subPath) {
        this.subPath = subPath;
    }

    public boolean isVisitProxy() {
        return isVisitProxy;
    }

    public void setVisitProxy(boolean isVisitProxy) {
        this.isVisitProxy = isVisitProxy;
    }

    public boolean isSetWatcher() {
        return isSetWatcher;
    }

    public void setSetWatcher(boolean isSetWatcher) {
        this.isSetWatcher = isSetWatcher;
    }

    public Map<String, String> getNotLabels() {
        return notLabels;
    }

    public void setNotLabels(Map<String, String> notLabels) {
        this.notLabels = notLabels;
    }

    public Map<String, List<String>> getInLabels() {
        return inLabels;
    }

    public void setInLabels(Map<String, List<String>> inLabels) {

```

```
        this.inLabels = inLabels;
    }

    public Map<String, List<String>> getNotInLabels() {
        return notInLabels;
    }

    public void setNotInLabels(Map<String, List<String>> notInLabels) {
        this.notInLabels = notInLabels;
    }
}
```

Params 对象包含的属性说明如表 4.2 所示。

表 4.2 Params 对象包含的属性说明

属 性	说 明
namespace	String 类型属性，指明资源所在的命名空间，如果没有指定该值，则表明访问所有命名空间下的资源对象
name	String 类型属性，在访问单个资源对象时使用，如果没有指定该值，则表明访问该类资源列表
fields	Map<String, String>类型属性，通过资源对象的域值过滤访问结果
labels	Map<String, String>类型属性，通过指定的标签选择器列表来选择资源对象。选择出的资源对象包含标签列表中所列的标签（即 Map 的 key），且所选资源的标签的 value 和标签列表中的 value 值（即 Map 的 value）相等
notLabels	Map<String, String>类型属性，通过指定的标签选择器列表来选择资源对象。选择出的资源对象包含标签列表中所列的标签（即 Map 的 key），且所选资源的标签的 value 和标签列表中的 value 值（即 Map 的 value）不相等
inLabels	Map<String, List<String>>类型属性，通过指定的标签选择器列表来选择资源对象。Map 对象的 key 值为标签名称，Map 对象的 value 值为该标签可能包含的值
notInLabels	Map<String, List<String>>类型属性，通过指定的标签选择器列表来选择资源对象。Map 对象的 key 值为标签名称，Map 对象的 value 值为列表，表明资源对象包含和 key 值同名的标签，且这些标签的值不在该列表中
json	String 类型属性，在创建或修改资源对象时使用，用于向 API Server 提供资源对象的定义
resourceType	ResourceType 类型属性，用于指明访问资源对象的类型
subPath	String 类型属性，用于指明访问资源的子目录
isVisitProxy	Boolean 类型属性，用于指明是否通过 Proxy 的方式访问资源对象
isSetWatcher	Boolean 类型属性，表明是否通过 Watcher 方式访问资源对象

Params 的 buildPath 方法用于构建访问 URL 的完整路径。

接口对象 RestfulClient 定义了访问 API 接口的所有方法（Method），其代码列表如下：

```
package com.hp.k8s.apiclient;

import com.hp.k8s.apiclient.imp.Params;

public interface RestfulClient {
    public String get(Params params); //获得单个资源对象
    public String list(Params params); //获得资源对象列表
    public String create(Params params); //创建资源对象
}
```

```

        public String delete(Params params); //删除某个资源对象
        public String update(Params params); //部分更新某个资源对象
        public String updateWithMediaType(Params params,String mediaType); //通过
mediaType, 实现 Merge
        public String replace(Params params); //替换某个资源对象
        public String options(Params params);
        public String head(Params params);
    }
}

```

其中 `get` 和 `list` 方法对应 Kubernetes API 的 `GET` 方法; `create` 方法对应 API 中的 `POST` 方法; `delete` 方法对应 API 中的 `DELETE` 方法; `update` 方法对应 API 中的 `PATCH` 方法; `replace` 方法对应 API 中的 `PUT` 方法; `options` 方法对应 API 中的 `OPTIONS` 方法; `head` 方法对应 API 中的 `HEAD` 方法。

该接口的基于 Jersey 框架的实现类如下所示:

```

package com.hp.k8s.apiclient.imp;

import javax.ws.rs.core.MediaType;

import org.apache.logging.log4j.LogManager;
import org.apache.logging.log4j.Logger;

import com.hp.k8s.apiclient.RestfulClient;
import com.sun.jersey.api.client.Client;
import com.sun.jersey.api.client.WebResource;
import com.sun.jersey.api.client.config.DefaultClientConfig;
import com.sun.jersey.client.urlconnection.URLConnectionClientHandler;

public class JerseyRestfulClient implements RestfulClient {
    private static final Logger LOG = LogManager.getLogger(RestfulClient.
class.getName());
    private static final String METHOD_PATCH = "PATCH";

    private String _baseUrl = null;
    Client _client = null;

    public JerseyRestfulClient(String baseUrl) {
        DefaultClientConfig config = new DefaultClientConfig();
        config.getProperties().put(URLConnectionClientHandler.PROPERTY_HTTP_
URL_CONNECTION_SET_METHOD_WORKAROUND, true);
        _client = Client.create(config);

        this._baseUrl = baseUrl;
    }
}

```

```
@Override
public String get(Params params) {
    WebResource resource = _client.resource(_baseUrl + params.buildPath());
    String response = resource.accept(MediaType.APPLICATION_JSON_TYPE).
get(String.class);
    LOG.info("Get one resource:\n" + response);

    return response;
}

@Override
public String list(Params params) {
    WebResource resource = _client.resource(_baseUrl + params.buildPath());
    LOG.info("URL: " + _baseUrl + params.buildPath());
    String response = resource.accept(MediaType.APPLICATION_JSON_TYPE).
get(String.class);

    return response;
}

@Override
public String create(Params params) {
    WebResource resource = _client.resource(_baseUrl + params.buildPath());
    LOG.info("URL: " + _baseUrl + params.buildPath());
    LOG.info("Create resource: " + params.getJson());
    String response = (null == params.getJson())
        ? resource.accept(MediaType.APPLICATION_JSON).post(String.class)
        : resource.type(MediaType.APPLICATION_JSON).accept(MediaType.
APPLICATION_JSON).post(String.class,
        params.getJson());

    return response;
}

@Override
public String delete(Params params) {
    WebResource resource = _client.resource(_baseUrl + params.buildPath());
    String response = resource.accept(MediaType.APPLICATION_JSON_TYPE).
delete(String.class);
    LOG.info("Delete resource " + params.getResourceType().getType() + "/"
+ params.getName() + " result:\n"
        + response);

    return response;
}

@Override
```

```

public String update(Params params) {
    return updateWithMediaType(params, MediaType.APPLICATION_JSON);
}

@Override
public String updateWithMediaType(Params params, String mediaType) {
    WebResource resource = _client.resource(_baseUrl + params.buildPath());
    LOG.info("URL: " + _baseUrl + params.buildPath());
    LOG.info("Patch resource: " + params.getJson());
    String response = resource.type(mediaType).accept(MediaType.APPLICATION_
JSON_TYPE).method(METHOD_PATCH, String.class,
        params.getJson());
    LOG.info("Update resource " + params.buildPath() + " result:\n" + response);

    return response;
}

@Override
public String replace(Params params) {
    WebResource resource = _client.resource(_baseUrl + params.buildPath());
    LOG.info("URL: " + _baseUrl + params.buildPath());
    LOG.info("Replace resource: " + params.getJson());
    String response = resource.type(MediaType.APPLICATION_JSON_TYPE).accept
(MediaType.APPLICATION_JSON_TYPE)
        .put(String.class, params.getJson());
    LOG.info("Replace resource " + params.buildPath() + " result:\n" + response);

    return response;
}

@Override
public String options(Params params) {
    WebResource resource = _client.resource(_baseUrl + params.buildPath());
    String response = resource.type(MediaType.APPLICATION_JSON_TYPE).accept
(MediaType.TEXT_PLAIN_TYPE)
        .options(String.class);
    LOG.info("Get options for resource " + params.getResourceType().getType()
+ "/" + params.getName()
        + " result:\n" + response);

    return response;
}

@Override
public String head(Params params) {
    WebResource resource = _client.resource(_baseUrl + params.buildPath());
    String response = resource.accept(MediaType.TEXT_PLAIN_TYPE).head().

```

```
getResponseStatus().toString();
    LOG.info("Get head for resource " + params.getResourceType().getType() +
"/" + params.getName() + " result:\n"
    + response);

    return response;
}

@Override
public void close() {
    _client.destroy();
}
}
```

该对象中包含如下代码：

```
config.getProperties().put(URLConnectionClientHandler.PROPERTY_HTTP_URL_CONNECTION_SET_METHOD_WORKAROUND, true);
```

该段代码的作用是使 Jersey 客户端能够支持除标准 REST 方法外的方法，比如 PATCH 方法。该段代码能访问除 watcher 外的所有 Kubernetes API 接口，在后续的章节中我们会举例说明如何访问 Kubernetes API。

4.3.2 Fabric8

Fabric8 包含多款工具包，Kubernetes Client 只是其中之一，也是 Kubernetes 官网中提到的 Java Client API 之一。本例代码涉及的 Jar 包如图 4.9 所示。




















 dnsjava-2.1.7.jar	2015/8/31 14:23	Executable Jar File	301 KB
 fabric8-utils-2.2.22.jar	2015/8/31 14:23	Executable Jar File	134 KB
 jackson-annotations-2.6.0.jar	2015/8/31 16:27	Executable Jar File	46 KB
 jackson-core-2.6.1.jar	2015/8/31 16:28	Executable Jar File	253 KB
 jackson-databind-2.6.1.jar	2015/8/31 15:56	Executable Jar File	1,140 KB
 jackson-dataformat-yaml-2.6.1.jar	2015/8/31 15:56	Executable Jar File	313 KB
 jackson-module-jaxb-annotations-2.6.0.jar	2015/8/31 16:24	Executable Jar File	32 KB
 json-20141113.jar	2015/8/31 14:23	Executable Jar File	64 KB
 kubernetes-api-2.2.22.jar	2015/8/31 14:22	Executable Jar File	72 KB
 kubernetes-client-1.3.8.jar	2015/8/31 15:37	Executable Jar File	2,262 KB
 kubernetes-model-1.0.12.jar	2015/8/31 15:56	Executable Jar File	2,308 KB
 log4j-api-2.3.jar	2015/8/31 16:18	Executable Jar File	133 KB
 log4j-core-2.3.jar	2015/8/31 15:56	Executable Jar File	808 KB
 log4j-slf4j-impl-2.3.jar	2015/8/31 15:56	Executable Jar File	23 KB
 oauth-20100527.jar	2015/8/31 15:56	Executable Jar File	44 KB
 openshift-client-1.3.2.jar	2015/8/31 14:23	Executable Jar File	24 KB
 slf4j-api-1.7.12.jar	2015/8/31 15:56	Executable Jar File	32 KB
 sundr-annotations-0.0.25.jar	2015/8/31 15:56	Executable Jar File	146 KB
 validation-api-1.1.0.Final.jar	2015/8/31 14:23	Executable Jar File	63 KB

图 4.9 本例代码涉及的 Jar 包

因为该工具包已经对访问 Kubernetes API 客户端做了较好的封装，因此其访问代码比较简单，其具体的访问过程会在后续的章节举例说明。

Fabric 8 的 Kubernetes API 客户端工具包只能访问 Node、Service、Pod、Endpoints、Events、Namespace、PersistentVolumeClaims、PersistentVolume、ReplicationController、ResourceQuota、Secret 和 ServiceAccount 这几个资源类型，不能使用 OPTIONS 和 HEAD 方法访问资源，且不能以代理方式访问资源，但其对以 watcher 方式访问资源做了很好的支持。

### 4.3.3 使用说明

首先，举例说明对 API 资源的基本访问，也就是对资源的增、删、改、查，以及替换资源的 status。其中会单独对 Node 和 Pod 的特殊接口做举例说明。表 4.3 列出了各资源对象的基本 API 接口。

表 4.3 各资源对象的基本 API 接口

资源类型	方法	URL Path	说明	备注
NODES	GET	/api/v1/nodes	获取 Node 列表	
	POST	/api/v1/nodes	创建一个 Node 对象	
	DELETE	/api/v1/nodes/{name}	删除一个 Node 对象	
	GET	/api/v1/nodes/{name}	获取一个 Node 对象	
	PATCH	/api/v1/nodes/{name}	部分更新一个 Node 对象	
	PUT	/api/v1/nodes/{name}	替换一个 Node 对象	
NAMESPACES	GET	/api/v1/namespaces	获取 Namespace 列表	
	POST	/api/v1/namespaces	创建一个 Namespace 对象	
	DELETE	/api/v1/namespaces/{name}	删除一个 Namespace 对象	
	GET	/api/v1/namespaces/{name}	获取一个 Namespace 对象	
	PATCH	/api/v1/namespaces/{name}	部分更新一个 Namespace 对象	
	PUT	/api/v1/namespaces/{name}	替换一个 Namespace 对象	
	PUT	/api/v1/namespaces/{name}/finalize	替换一个 Namespace 对象的最终方案对象	在 Fabric8 中没有实现
	PUT	/api/v1/namespaces/{name}/status	替换一个 Namespace 对象的状态	在 Fabric8 中没有实现
SERVICES	GET	/api/v1/services	获取 Service 列表	
	POST	/api/v1/services	创建一个 Service 对象	
	GET	/api/v1/namespaces/{namespace}/services	获取某个 Namespace 下的 Service 列表	
	POST	/api/v1/namespaces/{namespace}/services	在某个 Namespace 下创建列表	

续表

资源类型	方法	URL Path	说明	备注
	DELETE	/api/v1/namespaces/{namespace}/services/{name}	删除某个 Namespace 下的一个 Service 对象	
	GET	/api/v1/namespaces/{namespace}/services/{name}	获取某个 Namespace 下的一个 Service 对象	
	PATCH	/api/v1/namespaces/{namespace}/services/{name}	部分更新某个 Namespace 下的一个 Service 对象	
	PUT	/api/v1/namespaces/{namespace}/services/{name}	替换某个 Namespace 下的一个 Service 对象	
REPLICATIONCONTROLLERS	GET	/api/v1/replicationcontrollers	获取 RC 列表	
	POST	/api/v1/replicationcontrollers	创建一个 RC 对象	
	GET	/api/v1/namespaces/{namespace}/replicationcontrollers	获取某个 Namespace 下的 RC 列表	
	POST	/api/v1/namespaces/{namespace}/replicationcontrollers	在某个 Namespace 下创建一个 RC 对象	
	DELETE	/api/v1/namespaces/{namespace}/replicationcontrollers/{name}	删除某个 Namespace 下的 RC 对象	
	GET	/api/v1/namespaces/{namespace}/replicationcontrollers/{name}	获取某个 Namespace 下的 RC 对象	
	PATCH	/api/v1/namespaces/{namespace}/replicationcontrollers/{name}	部分更新某个 Namespace 下的 RC 对象	
	PUT	/api/v1/namespaces/{namespace}/replicationcontrollers/{name}	替换某个 Namespace 下的 RC 对象	
PODS	GET	/api/v1/pods	获取一个 Pod 列表	
	POST	/api/v1/pods	创建一个 Pod 对象	
	GET	/api/v1/namespaces/{namespace}/pods	获取某个 Namespace 下的 Pod 列表	
	POST	/api/v1/namespaces/{namespace}/pods	在某个 Namespace 下创建一个 Pod 对象	
	DELETE	/api/v1/namespaces/{namespace}/pods/{name}	删除某个 Namespace 下的一个 Pod 对象	
	GET	/api/v1/namespaces/{namespace}/pods/{name}	获取某个 Namespace 下的一个 Pod 对象	
	PATCH	/api/v1/namespaces/{namespace}/pods/{name}	部分更新某个 Namespace 下的一个 Pod 对象	
	PUT	/api/v1/namespaces/{namespace}/pods/{name}	替换某个 Namespace 下的一个 Pod 对象	



续表

资源类型	方法	URL Path	说明	备注
	PUT	/api/v1/namespaces/{namespace}/pods/{name}/status	替换某个 Namespace 下的一个 Pod 对象状态	在 Fabric8 中没有实现
	POST	/api/v1/namespaces/{namespace}/pods/{name}/binding	创建某个 Namespace 下的一个 Pod 对象的 Binding	在 Fabric8 中没有实现
	GET	/api/v1/namespaces/{namespace}/pods/{name}/exec	连接到某个 Namespace 下的一个 Pod 对象，并执行 exec	在 Fabric8 中没有实现
	POST	/api/v1/namespaces/{namespace}/pods/{name}/exec	连接到某个 Namespace 下的一个 Pod 对象，并执行 exec	在 Fabric8 中没有实现
	GET	/api/v1/namespaces/{namespace}/pods/{name}/log	连接到某个 Namespace 下的一个 Pod 对象，并获取 log 日志信息	在 Fabric8 中没有实现
	GET	/api/v1/namespaces/{namespace}/pods/{name}/portforward	连接到某个 Namespace 下的一个 Pod 对象，并实现端口转发	在 Fabric8 中没有实现
	POST	/api/v1/namespaces/{namespace}/pods/{name}/portforward	连接到某个 Namespace 下的一个 Pod 对象，并实现端口转发	在 Fabric8 中没有实现
BINDINGS	POST	/api/v1/bindings	创建一个 Binding 对象	
	POST	/api/v1/namespaces/{namespace}/bindings	在某个 Namespace 下创建一个 Binding 对象	
ENDPOINTS	GET	/api/v1/endpoints	获取 Endpoint 列表	
	POST	/api/v1/endpoints	创建一个 Endpoint 对象	
	GET	/api/v1/namespaces/{namespace}/endpoints	获取某个 Namespace 下的 Endpoint 对象列表	
	POST	/api/v1/namespaces/{namespace}/endpoints	在某个 Namespace 下创建一个 Endpoint 对象	
	DELETE	/api/v1/namespaces/{namespace}/endpoints/{name}	删除某个 Namespace 下的 Endpoint 对象	
	GET	/api/v1/namespaces/{namespace}/endpoints/{name}	获取某个 Namespace 下的 Endpoint 对象	
	PATCH	/api/v1/namespaces/{namespace}/endpoints/{name}	部分更新某个 Namespace 下的 Endpoint 对象	
	PUT	/api/v1/namespaces/{namespace}/endpoints/{name}	替换某个 Namespace 下的 Endpoint 对象	
SERVICEACCOUNTS	GET	/api/v1/serviceaccounts	获取 Serviceaccount 列表	
	POST	/api/v1/serviceaccounts	创建一个 Serviceaccount 对象	

续表

资源类型	方法	URL Path	说明	备注
	GET	/api/v1/namespaces/{namespace}/serviceaccounts	获取某个 Namespace 下的 Serviceaccount 对象列表	
	POST	/api/v1/namespaces/{namespace}/serviceaccounts	在某个 Namespace 下创建一个 Serviceaccount 对象	
	DELETE	/api/v1/namespaces/{namespace}/serviceaccounts/{name}	删除某个 Namespace 下的一个 Serviceaccount 对象	
	GET	/api/v1/namespaces/{namespace}/serviceaccounts/{name}	获取某个 Namespace 下的一个 Serviceaccount 对象	
	PATCH	/api/v1/namespaces/{namespace}/serviceaccounts/{name}	部分更新某个 Namespace 下的一个 Serviceaccount 对象	
	PUT	/api/v1/namespaces/{namespace}/serviceaccounts/{name}	替换某个 Namespace 下的一个 Serviceaccount 对象	
SECRETS	GET	/api/v1/secrets	获取 Secret 列表	
	POST	/api/v1/secrets	创建一个 Secret 对象	
	GET	/api/v1/namespaces/{namespace}/secrets	获取某个 Namespace 下的 Secret 列表	
	POST	/api/v1/namespaces/{namespace}/secrets	在某个 Namespace 下创建一个 Secret 对象	
	DELETE	/api/v1/namespaces/{namespace}/secrets/{name}	删除某个 Namespace 下的一个 Secret 对象	
	GET	/api/v1/namespaces/{namespace}/secrets/{name}	获取某个 Namespace 下的一个 Secret 对象	
	PATCH	/api/v1/namespaces/{namespace}/secrets/{name}	部分更新某个 Namespace 下的一个 Secret 对象	
	PUT	/api/v1/namespaces/{namespace}/secrets/{name}	替换某个 Namespace 下的一个 Secret 对象	
EVENTS	GET	/api/v1/events	获取 Event 列表	
	POST	/api/v1/events	创建一个 Event 对象	
	GET	/api/v1/namespaces/{namespace}/events	获取某个 Namespace 下的 Event 列表	
	POST	/api/v1/namespaces/{namespace}/events	在某个 Namespace 下创建一个 Event 对象	
	DELETE	/api/v1/namespaces/{namespace}/events/{name}	删除某个 Namespace 下的一个 Event 对象	
	GET	/api/v1/namespaces/{namespace}/events/{name}	获取某个 Namespace 下的一个 Event 对象	
	PATCH	/api/v1/namespaces/{namespace}/events/{name}	部分更新某个 Namespace 下的一个 Event 对象	

续表

资源类型	方法	URL Path	说明	备注
	PUT	/api/v1/namespaces/{namespace}/events/{name}	替换某个 Namespace 下的一个 Event 对象	
COMPONENTSTATUS	GET	/api/v1/componentstatuses	获取 ComponentStatus 列表	
	GET	/api/v1/namespaces/{namespace}/componentstatuses	获取某个 Namespace 下的 ComponentStatus 列表	
	GET	/api/v1/namespaces/{namespace}/componentstatuses/{name}	获取某个 Namespace 下的一个 ComponentStatus 对象	
LIMITRANGES	GET	/api/v1/limitranges	获取 LimitRange 列表	
	POST	/api/v1/limitranges	创建一个 LimitRange 对象	
	GET	/api/v1/namespaces/{namespace}/limitranges	获取某个 Namespace 下的 LimitRange 列表	
	POST	/api/v1/namespaces/{namespace}/limitranges	在某个 Namespace 下创建一个 LimitRange 对象	
	DELETE	/api/v1/namespaces/{namespace}/limitranges/{name}	删除某个 Namespace 下的一个 LimitRange 对象	
	GET	/api/v1/namespaces/{namespace}/limitranges/{name}	获取某个 Namespace 下的一个 LimitRange 对象	
	PATCH	/api/v1/namespaces/{namespace}/limitranges/{name}	部分更新某个 Namespace 下的一个 LimitRange 对象	
	PUT	/api/v1/namespaces/{namespace}/limitranges/{name}	替换某个 Namespace 下的一个 LimitRange 对象	
RESOURCEQUOTAS	GET	/api/v1/resourcequotas	获取 ResourceQuota 列表	
	POST	/api/v1/resourcequotas	创建一个 ResourceQuota 对象	
	GET	/api/v1/namespaces/{namespace}/resourcequotas	获取某个 Namespace 下的 ResourceQuota 列表	
	POST	/api/v1/namespaces/{namespace}/resourcequotas	在某个 Namespace 下创建一个 ResourceQuota 对象	
	DELETE	/api/v1/namespaces/{namespace}/resourcequotas/{name}	删除某个 Namespace 下的一个 ResourceQuota 对象	
	GET	/api/v1/namespaces/{namespace}/resourcequotas/{name}	获取某个 Namespace 下的一个 ResourceQuota 对象	
	PATCH	/api/v1/namespaces/{namespace}/resourcequotas/{name}	部分更新某个 Namespace 下的一个 ResourceQuota 对象	

续表

资源类型	方法	URL Path	说明	备注
	PUT	/api/v1/namespaces/{namespace}/resourcequotas/{name}	替换某个 Namespace 下的一个 Resource Quota 对象	
	PUT	/api/v1/namespaces/{namespace}/resourcequotas/{name}/status	替换某个 Namespace 下的一个 Resource Quota 对象状态	在 Fabric8 中没有实现
PODTEMPLATES	GET	/api/v1/podtemplates	获取 PodTemplate 列表	
	POST	/api/v1/podtemplates	创建一个 PodTemplate 对象	
	GET	/api/v1/namespaces/{namespace}/podtemplates	获取某个 Namespace 下的 PodTemplate 列表	
	POST	/api/v1/namespaces/{namespace}/podtemplates	在某个 Namespace 下创建一个 PodTemplate 对象	
	DELETE	/api/v1/namespaces/{namespace}/podtemplates/{name}	删除某个 Namespace 下的一个 PodTemplate 对象	
	GET	/api/v1/namespaces/{namespace}/podtemplates/{name}	获取某个 Namespace 下的一个 PodTemplate 对象	
	PATCH	/api/v1/namespaces/{namespace}/podtemplates/{name}	部分更新某个 Namespace 下的一个 PodTemplate 对象	
	PUT	/api/v1/namespaces/{namespace}/podtemplates/{name}	替换某个 Namespace 下的一个 PodTemplate 对象	
PERSISTENTVOLUMES	GET	/api/v1/persistentvolumes	获取 PersistentVolume 列表	
	POST	/api/v1/persistentvolumes	创建一个 PersistentVolume 对象	
	DELETE	/api/v1/persistentvolumes/{name}	删除一个 PersistentVolume 对象	
	GET	/api/v1/persistentvolumes/{name}	获取一个 PersistentVolume 对象	
	PATCH	/api/v1/persistentvolumes/{name}	部分更新一个 PersistentVolume 对象	
	PUT	/api/v1/persistentvolumes/{name}	替换一个 PersistentVolume 对象	
	PUT	/api/v1/persistentvolumes/{name}/status	替换一个 PersistentVolume 对象状态	在 Fabric8 中没有实现
PERSISTENTVOLUMECLAIMS	GET	/api/v1/persistentvolumeclaims	获取 PersistentVolumeClaim 列表	
	POST	/api/v1/persistentvolumeclaims	创建一个 PersistentVolumeClaim 对象	
	GET	/api/v1/namespaces/{namespace}/persistentvolumeclaims	获取某个 Namespace 下的 PersistentVolumeClaim 列表	
	POST	/api/v1/namespaces/{namespace}/persistentvolumeclaims	在某个 Namespace 下创建一个 PersistentVolumeClaim 对象	

续表

资源类型	方法	URL Path	说明	备注
	DELETE	/api/v1/namespaces/{namespace}/persistentvolumeclaims/{name}	删除某个 Namespace 下的一个 Persistent VolumeClaim 对象	
	GET	/api/v1/namespaces/{namespace}/persistentvolumeclaims/{name}	获取某个 Namespace 下的一个 Persistent VolumeClaim 对象	
	PATCH	/api/v1/namespaces/{namespace}/persistentvolumeclaims/{name}	部分更新某个 Namespace 下的一个 Persistent VolumeClaim 对象	
	PUT	/api/v1/namespaces/{namespace}/persistentvolumeclaims/{name}	替换某个 Namespace 下的一个 Persistent VolumeClaim 对象	
	PUT	/api/v1/namespaces/{namespace}/persistentvolumeclaims/{name}/status	替换某个 Namespace 下的一个 Persistent VolumeClaim 对象状态	在 Fabric8 中没有实现

首先，举例说明如何通过 API 接口来创建资源对象。我们需要创建访问 API Server 的客户端，基于 Jersey 框架的代码如下：

```
RestfulClient _restfulClient = new JerseyRestfulClient("http://192.168.1.128:8080/api/v1");
```

其中，http://192.168.1.128:8080 为 API Server 的地址。基于 Fabric8 框架的代码如下：

```
Config _conf = new Config();
KubernetesClient _kube = new DefaultKubernetesClient("http://192.168.1.128: 8080");
```

分别通过上面的两个客户端创建 Namespace 资源对象，基于 Jersey 框架的代码如下：

```
private void testCreateNamespace() {
    Params params = new Params();
    params.setResourceType(ResourceType.NAMESPACES);
    params.setJson(Utils.getJson("namespace.json"));

    LOG.info("Result: " + _restfulClient.create(params));
}
```

其中，“namespace.json”为创建 Namespace 资源对象的 JSON 定义，代码如下：

```
{
  "kind": "Namespace",
  "apiVersion": "v1",
  "metadata": {
    "name": "ns-sample"
  }
}
```

基于 Fabric8 框架的代码如下：

```
private void testCreateNamespace() {
    Namespace ns = new Namespace();
```

```

ns.setApiVersion(ApiVersion.V_1);
ns.setKind("Namespace");
ObjectMeta om = new ObjectMeta();
om.setName("ns-fabric8");
ns.setMetadata(om);

_kube.namespaces().create(ns);

LOG.info(_kube.namespaces().list().getItems().size());
}

```

由于 Fabric8 框架对 Kubernetes API 对象做了很好的封装，对其中的大量对象都做了定义，所以用户可以通过其提供的资源对象去定义 Kubernetes API 对象，例如上面例子中的 Namespace 对象。Fabric8 框架中的 `kubernetes-model` 工具包用于 API 对象的封装。在上面的例子中，通过 Fabric8 框架提供的类创建了一个名为“ns-fabric8”的命名空间对象。

接下来我们会通过基于 Jeysey 框架的代码去创建两个 Pod 资源对象。在两个例子中，一个是在上面创建的“ns-sample”Namespace 中创建 Pod 资源对象，另一个是为后续创建“cluster service”而创建的 Pod 资源对象。由于基于 Fabric8 框架创建 Pod 资源对象的方法很简单，因此不再用 Fabric8 框架对上述两个例子做说明。通过基于 Jersey 框架创建这两个 Pod 资源对象的代码如下：

```

private void testCreatePod() {
    Params params = new Params();
    params.setResourceType(ResourceType.PODS);
    params.setJson(Utils.getJson("podInNs.json"));
    params.setNamespace("ns-sample");
    LOG.info("Result: " + _restfulClient.create(params));

    params.setJson(Utils.getJson("pod4ClusterService.json"));
    LOG.info("Result: " + _restfulClient.create(params));
}

```

其中，`podInNs.json` 和 `pod4ClusterService.json` 是创建两个 Pod 资源对象的定义。`podInNs.json` 文件的内容如下：

```

{
  "kind": "Pod",
  "apiVersion": "v1",
  "metadata": {
    "name": "pod-sample-in-namespace",
    "namespace": "ns-sample"
  },
  "spec": {
    "containers": [{
      "name": "mycontainer",

```

```

        "image": "kubeguide/redis-master"
    }]
}
}

```

pod4ClusterService.json 文件的内容如下:

```

{
  "kind": "Pod",
  "apiVersion": "v1",
  "metadata": {
    "name": "pod-sample-4-cluster-service",
    "namespace": "ns-sample",
    "labels": {
      "k8s-cs": "kube-cluster-service",
      "k8s-test": "kube-cluster-test",
      "k8s-sample-app": "kube-service-sample",
      "kkk": "bbb"
    }
  },
  "spec": {
    "containers": [{
      "name": "mycontainer",
      "image": "kubeguide/redis-master"
    }]
  }
}

```

下面的例子代码用于获取 Pod 资源列表,其中第 1 部分代码用于获取所有的 Pod 资源对象,第 2、3 部分代码主要是列举如何使用标签选择 Pod 资源对象,最后一部分代码用于举例说明如何使用 field 选择 Pod 资源对象。代码如下:

```

private void testGetPodList() {
    Params params = new Params();
    params.setResourceType(ResourceType.PODS);
    LOG.info("Result: " + _restfulClient.list(params));

    Map<String, String> labels = new HashMap<String, String>();
    labels.put("k8s-cs", "kube-cluster-service");
    labels.put("k8s-sample-app", "kube-service-sample");
    params.setLabels(labels);
    LOG.info("Result: " + _restfulClient.list(params));
    params.setLabels(null);

    Map<String, List<String>> inLabels = new HashMap<String, List<String>>();
    List list = new ArrayList<String>();
    list.add("kube-cluster-service");
    list.add("kube-cluster");
}

```

```
inLabels.put("k8s-cs", list);
params.setInLabels(inLabels);
LOG.info("Result: " + _restfulClient.list(params));
params.setInLabels(null);

Map<String, String> fields = new HashMap<String, String>();
fields.put("metadata.name", "pod-sample-4-cluster-service");
params.setNamespace("ns-sample");
params.setFields(fields);
LOG.info("Result: " + _restfulClient.list(params));
}
```

接下来的例子代码用于替换一个 Pod 对象，在通过 Kubernetes API 替换一个 Pod 资源对象时需要注意如下两点。

(1) 在替换该资源对象前，先从 API 中获取该资源对象的 JSON 对象，然后在该 JSON 对象的基础上修改需要替换的部分。

(2) 在 Kubernetes API 提供的接口中，PUT 方法（replace）只支持替换容器的 image 部分。

代码如下：

```
private void testReplacePod() {
    Params params = new Params();
    params.setNamespace("ns-sample");
    params.setName("pod-sample-in-namespace");
    params.setJson(Utils.getJson("pod4Replace.json"));
    params.setResourceType(ResourceType.PODS);

    LOG.info("Result: " + _restfulClient.replace(params));
}
```

其中，pod4Replace.json 的内容如下：

```
{
  "kind": "Pod",
  "apiVersion": "v1",
  "metadata": {
    "name": "pod-sample-in-namespace",
    "namespace": "ns-sample",
    "selfLink": "/api/v1/namespaces/ns-sample/pods/pod-sample-in-namespace",
    "uid": "084ff63e-59d3-11e5-8035-000c2921ba71",
    "resourceVersion": "45450",
    "creationTimestamp": "2015-09-13T04:51:01Z"
  },
  "spec": {
    "volumes": [
      {
        "name": "default-token-szoje",
```



```

    "secret": {
      "secretName": "default-token-szoje"
    }
  ],
  "containers": [
    {
      "name": "mycontainer",
      "image": "centos",
      "resources": {},
      "volumeMounts": [
        {
          "name": "default-token-szoje",
          "readOnly": true,
          "mountPath": "/var/run/secrets/kubernetes.io/serviceaccount"
        }
      ],
      "terminationMessagePath": "/dev/termination-log",
      "imagePullPolicy": "IfNotPresent"
    }
  ],
  "restartPolicy": "Always",
  "dnsPolicy": "ClusterFirst",
  "serviceAccountName": "default",
  "serviceAccount": "default",
  "nodeName": "192.168.1.129"
},
"status": {
  "phase": "Running",
  "conditions": [
    {
      "type": "Ready",
      "status": "True"
    }
  ],
  "hostIP": "192.168.1.129",
  "podIP": "10.1.10.66",
  "startTime": "2015-09-11T15:17:28Z",
  "containerStatuses": [
    {
      "name": "mycontainer",
      "state": {
        "running": {
          "startedAt": "2015-09-11T15:17:30Z"
        }
      },
      "lastState": {},
    }
  ],

```

```
        "ready": true,
        "restartCount": 0,
        "image": "kubeguide/redis-master",
        "imageID":
"docker://5630952871a38cddffda9ec611f5978ab0933628fcd54cd7d7677ce6b17de33f",
        "containerID": "docker://7bf0d454c367418348711556e667fd1ef6a04d7153d
24bfcac2e2e06da634a9f"
    }
}
}
```

接下来的两个例子实现了 4.2.4 节中提到的两种 Merge 方式：Merge Patch 和 Strategic Merge Patch。

第 1 种 Merge 方式的示例如下：

```
private void testUpdatePod1() {
    Params params = new Params();
    params.setNamespace("ns-sample");
    params.setName("pod-sample-in-namespace");
    params.setJson(Utils.getJson("pod4MergeJsonPatch.json"));
    params.setResourceType(ResourceType.PODS);

    LOG.info("Result: " + _restfulClient.updateWithMediaType(params,
"application/ merge-patch+json"));
}
```

其中，pod4MergeJsonPatch.json 的内容如下：

```
{
  "metadata":{
    "labels":{
      "k8s-cs": "kube-cluster-service",
      "k8s-test": "kube-cluster-test",
      "k8s-sa555mple-app": "kube-service-sample",
      "kkk": "bbb4444"
    }
  }
}
```

第 2 种 Merge 方式（Strategic Merge Patch）的示例如下：

```
private void testUpdatePod2() {
    Params params = new Params();
    params.setNamespace("ns-sample");
    params.setName("pod-sample-in-namespace");
    params.setJson(Utils.getJson("pod4StrategicMerge.json"));
    params.setResourceType(ResourceType.PODS);
}
```

```
LOG.info("Result: " + _restfulClient.updateWithMediaType(params,
"application/strategic-merge-patch+json"));
}
```

其中，pod4StrategicMerge.json 的内容如下：

```
{
  "spec": {
    "containers": [ {
      "name": "mycontainer",
      "image": "centos",
      "patchStrategy": "merge",
      "patchMergeKey": "name"
    } ]
  }
}
```

接下来实现了修改 Pod 资源对象的状态，代码如下：

```
private void testStatusPod() {
    Params params = new Params();
    params.setNamespace("ns-sample");
    params.setName("pod-sample-in-namespace");
    params.setSubPath("/status");
    params.setJson(Utils.getJson("pod4Status.json"));
    params.setResourceType(ResourceType.PODS);

    _restfulClient.replace(params);
}
```

其中，pod4Status.json 的内容如下：

```
{
  "kind": "Pod",
  "apiVersion": "v1",
  "metadata": {
    "name": "pod-sample-in-namespace",
    "namespace": "ns-sample",
    "selfLink": "/api/v1/namespaces/ns-sample/pods/pod-sample-in-namespace",
    "uid": "ad1d803f-59ec-11e5-8035-000c2921ba71",
    "resourceVersion": "51640",
    "creationTimestamp": "2015-09-13T07:54:35Z"
  },
  "spec": {
    "volumes": [
      {
        "name": "default-token-szoje",
        "secret": {
          "secretName": "default-token-szoje"
        }
      }
    ]
  }
}
```

```
    }
  }
],
"containers": [
  {
    "name": "mycontainer",
    "image": "kubeguide/redis-master",
    "resources": {},
    "volumeMounts": [
      {
        "name": "default-token-szoje",
        "readOnly": true,
        "mountPath": "/var/run/secrets/kubernetes.io/serviceaccount"
      }
    ],
    "terminationMessagePath": "/dev/termination-log",
    "imagePullPolicy": "IfNotPresent"
  }
],
"restartPolicy": "Always",
"dnsPolicy": "ClusterFirst",
"serviceAccountName": "default",
"serviceAccount": "default",
"nodeName": "192.168.1.129"
},
"status": {
  "phase": "Unknown",
  "conditions": [
    {
      "type": "Ready",
      "status": "false"
    }
  ]
},
"hostIP": "192.168.1.129",
"podIP": "10.1.10.79",
"startTime": "2015-09-11T18:21:02Z",
"containerStatuses": [
  {
    "name": "mycontainer",
    "state": {
      "running": {
        "startedAt": "2015-09-11T18:21:03Z"
      }
    },
    "lastState": {},
    "ready": true,
    "restartCount": 0,
```

```

        "image": "kubeguide/redis-master",
        "imageID": "docker://5630952871a38cddffda9ec611f5978ab0933628fcd54cd
7d7677ce6b17de33f",
        "containerID": "docker://b0e2312643e9a4b59cf1ff5fb7a8468c5777180d5a
8ea5f2f0c9dfddcf3f4cd2"
    }
}
}
}

```

接下来实现了查看 Pod 的 log 日志功能，代码如下：

```

private void testLogPod() {
    Params params = new Params();
    params.setNamespace("ns-sample");
    params.setName("pod-sample-in-namespace");
    params.setSubPath("/log");
    params.setResourceType(ResourceType.PODS);

    _restfulClient.get(params);
}

```

下面通过 API 访问 Node 的多种接口，代码如下：

```

private void testPoxyNode() {
    Params params = new Params();
    params.setName("192.168.1.129");
    params.setSubPath("pods");
    params.setVisitProxy(true);
    params.setResourceType(ResourceType.NODES);
    _restfulClient.get(params);

    params = new Params();
    params.setName("192.168.1.129");
    params.setSubPath("stats");
    params.setVisitProxy(true);
    params.setResourceType(ResourceType.NODES);
    _restfulClient.get(params);

    params = new Params();
    params.setName("192.168.1.129");
    params.setSubPath("spec");
    params.setVisitProxy(true);
    params.setResourceType(ResourceType.NODES);
    _restfulClient.get(params);

    params = new Params();
    params.setName("192.168.1.129");
    params.setSubPath("run/ns-sample/pod/pod-sample-in-namespace");
}

```

```
params.setVisitProxy(true);
params.setResourceType(ResourceType.NODES);
_restfulClient.get(params);

params = new Params();
params.setName("192.168.1.129");
params.setSubPath("metrics");
params.setVisitProxy(true);
params.setResourceType(ResourceType.NODES);
_restfulClient.get(params);
}
```

最后，举例说明如何通过 API 删除资源对象 pod，代码如下：

```
private void testDetetePod() {
    Params params = new Params();
    params.setNamespace("ns-sample");
    params.setName("pod-sample-in-namespace");
    params.setResourceType(ResourceType.PODS);
    LOG.info("Result: " + _restfulClient.delete(params));
}
```

通过 API 接口除了能够对资源对象实现前面列出的基本操作外，还涉及两类特殊接口，一类是 WATCH，一类是 PROXY。这两类特殊接口所包含的接口如表 4.4 所示。

表 4.4 两类特殊接口所包含的接口

资源类型	类别	方法	URL Path	说明
NODES	WATCH	GET	/api/v1/watch/nodes	监听所有节点的变化
		GET	/api/v1/watch/nodes/{name}	监听单个节点的变化
	PROXY	DELETE	/api/v1/proxy/nodes/{name}/{path:~}	代理 DELETE 请求到节点的某个子目录
		GET	/api/v1/proxy/nodes/{name}/{path:~}	代理 GET 请求到节点的某个子目录
		HEAD	/api/v1/proxy/nodes/{name}/{path:~}	代理 HEAD 请求到节点的某个子目录
		OPTIONS	/api/v1/proxy/nodes/{name}/{path:~}	代理 OPTIONS 请求到节点的某个子目录
		POST	/api/v1/proxy/nodes/{name}/{path:~}	代理 POST 请求到节点的某个子目录
		PUT	/api/v1/proxy/nodes/{name}/{path:~}	代理 PUT 请求到节点的某个子目录
		DELETE	/api/v1/proxy/nodes/{name}	代理 DELETE 请求到节点
		GET	/api/v1/proxy/nodes/{name}	代理 GET 请求到节点
		HEAD	/api/v1/proxy/nodes/{name}	代理 HEAD 请求到节点
		OPTIONS	/api/v1/proxy/nodes/{name}	代理 OPTIONS 请求到节点
		POST	/api/v1/proxy/nodes/{name}	代理 POST 请求到节点
		PUT	/api/v1/proxy/nodes/{name}	代理 PUT 请求到节点

续表

资 源 类 型	类 别	方 法	URL Path	说 明
SERVICES	WATCH	GET	/api/v1/watch/services	监听所有 Service 的变化
		GET	/api/v1/watch/namespaces/{namespace}/services	监听某个 Namespace 下所有 Service 的变化
	PROXY	GET	/api/v1/watch/namespaces/{namespace}/services/{name}	监听某个 Service 的变化
		DELETE	/api/v1/proxy/namespaces/{namespace}/services/{name}/{path:~}	代理 DELETE 请求到 Service 的某个子目录
		GET	/api/v1/proxy/namespaces/{namespace}/services/{name}/{path:~}	代理 GET 请求到 Service 的某个子目录
		HEAD	/api/v1/proxy/namespaces/{namespace}/services/{name}/{path:~}	代理 HEAD 请求到 Service 的某个子目录
		OPTIONS	/api/v1/proxy/namespaces/{namespace}/services/{name}/{path:~}	代理 OPTIONS 请求到 Service 的某个子目录
		POST	/api/v1/proxy/namespaces/{namespace}/services/{name}/{path:~}	代理 POST 请求到 Service 的某个子目录
		PUT	/api/v1/proxy/namespaces/{namespace}/services/{name}/{path:~}	代理 PUT 请求到 Service 的某个子目录
		DELETE	/api/v1/proxy/namespaces/{namespace}/services/{name}	代理 DELETE 请求到 Service
		GET	/api/v1/proxy/namespaces/{namespace}/services/{name}	代理 GET 请求到 Service
		HEAD	/api/v1/proxy/namespaces/{namespace}/services/{name}	代理 HEAD 请求到 Service
		OPTIONS	/api/v1/proxy/namespaces/{namespace}/services/{name}	代理 OPTIONS 请求到 Service
		POST	/api/v1/proxy/namespaces/{namespace}/services/{name}	代理 POST 请求到 Service
		PUT	/api/v1/proxy/namespaces/{namespace}/services/{name}	代理 PUT 请求到 Service
REPLICATIONCONTROLLER	WATCH	GET	/api/v1/watch/replicationcontrollers	监听所有 RC 的变化
		GET	/api/v1/watch/namespaces/{namespace}/replicationcontrollers	监听某个 Namespace 下所有 RC 的变化
		GET	/api/v1/watch/namespaces/{namespace}/replicationcontrollers/{name}	监听某个 RC 的变化

续表

资源类型	类别	方法	URL Path	说明
PODS	WATCH	GET	/api/v1/watch/pods	监听所有 Pod 的变化
		GET	/api/v1/watch/namespaces/{namespace}/pods	监听某个 Namespace 下所有 Pod 的变化
		GET	/api/v1/watch/namespaces/{namespace}/pods/{name}	监听某个 Pod 的变化
	PROXY	DELETE	/api/v1/namespaces/{namespace}/pods/{name}/proxy/{path:*}	代理 DELETE 请求到 Pod 的某个子目录
		GET	/api/v1/namespaces/{namespace}/pods/{name}/proxy/{path:*}	代理 GET 请求到 Pod 的某个子目录
		HEAD	/api/v1/namespaces/{namespace}/pods/{name}/proxy/{path:*}	代理 HEAD 请求到 Pod 的某个子目录
		OPTIONS	/api/v1/namespaces/{namespace}/pods/{name}/proxy/{path:*}	代理 OPTIONS 请求到 Pod 的某个子目录
		POST	/api/v1/namespaces/{namespace}/pods/{name}/proxy/{path:*}	代理 POST 请求到 Pod 的某个子目录
		PUT	/api/v1/namespaces/{namespace}/pods/{name}/proxy/{path:*}	代理 PUT 请求到 Pod 的某个子目录
		DELETE	/api/v1/namespaces/{namespace}/pods/{name}/proxy	代理 DELETE 请求到 Pod
		GET	/api/v1/namespaces/{namespace}/pods/{name}/proxy	代理 GET 请求到 Pod
		HEAD	/api/v1/namespaces/{namespace}/pods/{name}/proxy	代理 HEAD 请求到 Pod
		OPTIONS	/api/v1/namespaces/{namespace}/pods/{name}/proxy	代理 OPTIONS 请求到 Pod
		POST	/api/v1/namespaces/{namespace}/pods/{name}/proxy	代理 POST 请求到 Pod
		PUT	/api/v1/namespaces/{namespace}/pods/{name}/proxy	代理 PUT 请求到 Pod
		DELETE	/api/v1/proxy/namespaces/{namespace}/pods/{name}/{path:*}	代理 DELETE 请求到 Pod 的某个子目录
		GET	/api/v1/proxy/namespaces/{namespace}/pods/{name}/{path:*}	代理 GET 请求到 Pod 的某个子目录
		HEAD	/api/v1/proxy/namespaces/{namespace}/pods/{name}/{path:*}	代理 HEAD 请求到 Pod 的某个子目录
		OPTIONS	/api/v1/proxy/namespaces/{namespace}/pods/{name}/{path:*}	代理 OPTIONS 请求到 Pod 的某个子目录
		POST	/api/v1/proxy/namespaces/{namespace}/pods/{name}/{path:*}	代理 POST 请求到 Pod 的某个子目录



续表

资源类型	类别	方法	URL Path	说明
		PUT	/api/v1/proxy/namespaces/{namespace}/pods/{name}/{path:*}	代理 PUT 请求到 Pod 的某个子目录
		DELETE	/api/v1/proxy/namespaces/{namespace}/pods/{name}	代理 DELETE 请求到 Pod
		GET	/api/v1/proxy/namespaces/{namespace}/pods/{name}	代理 GET 请求到 Pod
		HEAD	/api/v1/proxy/namespaces/{namespace}/pods/{name}	代理 HEAD 请求到 Pod
		OPTIONS	/api/v1/proxy/namespaces/{namespace}/pods/{name}	代理 OPTIONS 请求到 Pod
		POST	/api/v1/proxy/namespaces/{namespace}/pods/{name}	代理 POST 请求到 Pod
		PUT	/api/v1/proxy/namespaces/{namespace}/pods/{name}	代理 PUT 请求到 Pod
ENDPOINTS	WATCH	GET	/api/v1/watch/endpoints	监听所有 Endpoint 的变化
		GET	/api/v1/watch/namespaces/{namespace}/endpoints	监听某个 Namespace 下所有 Endpoint 的变化
		GET	/api/v1/watch/namespaces/{namespace}/endpoints/{name}	监听某个 Endpoint 的变化
SERVICEACCOUNT	WATCH	GET	/api/v1/watch/serviceaccounts	监听所有 ServiceAccount 的变化
		GET	/api/v1/watch/namespaces/{namespace}/serviceaccounts	监听某个 Namespace 下所有 ServiceAccount 的变化
		GET	/api/v1/watch/namespaces/{namespace}/serviceaccounts/{name}	监听某个 ServiceAccount 的变化
SECRET	WATCH	GET	/api/v1/watch/secrets	监听所有 Secret 的变化
		GET	/api/v1/watch/namespaces/{namespace}/secrets	监听某个 Namespace 下所有 Secret 的变化
		GET	/api/v1/watch/namespaces/{namespace}/secrets/{name}	监听某个 Secret 的变化
EVENTS	WATCH	GET	/api/v1/watch/events	监听所有 Event 的变化
		GET	/api/v1/watch/namespaces/{namespace}/events	监听某个 Namespace 下所有 Event 的变化

续表

资源类型	类别	方法	URL Path	说明
		GET	/api/v1/watch/namespaces/{namespace}/events/{name}	监听某个 Event 的变化
LIMITRANGES	WATCH	GET	/api/v1/watch/limitranges	监听所有 Event 的变化
		GET	/api/v1/watch/namespaces/{namespace}/limitranges	监听某个 Namespace 下所有 Event 的变化
		GET	/api/v1/watch/namespaces/{namespace}/limitranges/{name}	监听某个 Event 的变化
RESOURCEQUOTAS	WATCH	GET	/api/v1/watch/resourcequotas	监听所有 ResourceQuota 的变化
		GET	/api/v1/watch/namespaces/{namespace}/resourcequotas	监听某个 Namespace 下所有 ResourceQuota 的变化
		GET	/api/v1/watch/namespaces/{namespace}/resourcequotas/{name}	监听某个 ResourceQuota 的变化
PODTEMPLATES	WATCH	GET	/api/v1/watch/podtemplates	监听所有 PodTemplate 的变化
		GET	/api/v1/watch/namespaces/{namespace}/podtemplates	监听某个 Namespace 下所有 PodTemplate 的变化
		GET	/api/v1/watch/namespaces/{namespace}/podtemplates/{name}	监听某个 PodTemplate 的变化
PERSISTENTVOLUMES	WATCH	GET	/api/v1/watch/persistentvolumes	监听所有 PersistentVolume 的变化
		GET	/api/v1/watch/persistentvolumes/{name}	监听某个 PersistentVolume 的变化
PERSISTENTVOLUMECLAIMS	WATCH	GET	/api/v1/watch/persistentvolumeclaims	监听所有 PersistentVolumeClaim 的变化
		GET	/api/v1/watch/namespaces/{namespace}/persistentvolumeclaims	监听某个 Namespace 下所有 PersistentVolumeClaim 的变化
		GET	/api/v1/watch/namespaces/{namespace}/persistentvolumeclaims/{name}	监听某个 PersistentVolumeClaim 的变化

下面基于 Fabric8 实现对资源对象的监听（Watch），代码如下：

```
private void testWatcher() {
    _kube.pods().watch(new io.fabric8.kubernetes.client.Watcher<Pod>() {
        @Override
        public void eventReceived(Action action, Pod pod) {
            System.out.println(action + ": " + pod);
        }
    })
}
```

```

        @Override
        public void onClose(KubernetesClientException e) {
            System.out.println("Closed: " + e);
        }
    });
}

```

接下来基于 Jersey 框架实现通过 Proxy 方式访问 Pod。由于 API Server 针对 Pod 资源提供了两种 Proxy 访问接口，所以下面分别用两段代码进行示例说明。代码如下：

```

private void testPoxyPod() {
    //访问第1种 proxy 接口
    Params params = new Params();
    params.setNamespace("ns-sample");
    params.setName("pod-sample-in-namespace");
    params.setSubPath("/proxy");
    params.setResourceType(ResourceType.PODS);

    _restfulClient.get(params);

    //访问第2种 proxy 接口
    params = new Params();
    params.setNamespace("ns-sample");
    params.setName("pod-sample-in-namespace");
    params.setVisitProxy(true);
    params.setResourceType(ResourceType.PODS);

    _restfulClient.get(params);
}

```

# 第 5 章

## Kubernetes 运维指南

---

为了让容器应用在 Kubernetes 集群中运行得更加有效，对 Kubernetes 集群本身也需要进行相应的配置和管理。本章将从 Kubernetes 集群管理和 Trouble Shooting 等方面对 Kubernetes 集群的运维管理进行详细说明，最后对 Kubernetes 开发中的新功能进行介绍。

### 5.1 Kubernetes 集群管理指南

---

本节将从 Node 的管理、Label 的管理、Namespace 资源共享、资源配额管理、集群 Master 高可用及集群监控等方面，对 Kubernetes 集群本身的运维管理进行详细说明。

#### 5.1.1 Node 的管理

---

##### 1. Node 的隔离与恢复

在硬件升级、硬件维护等情况下，我们需要将某些 Node 进行隔离，脱离 Kubernetes 集群的调度范围。Kubernetes 提供了一种机制，既可以将 Node 纳入调度范围，也可以将 Node 脱离调度范围。

创建配置文件 `unschedule_node.yaml`，在 `spec` 部分指定 `unschedulable` 为 `true`：

```
apiVersion: v1
kind: Node
metadata:
  name: k8s-node-1
```

```
labels:
  kubernetes.io/hostname: k8s-node-1
spec:
  unschedulable: true
```

然后，通过 `kubectrl replace` 命令完成对 Node 状态的修改：

```
$ kubectrl replace -f unschedule_node.yaml
node "k8s-node-1" replaced
```

查看 Node 的状态，可以观察到在 Node 的状态中增加了一项 `SchedulingDisabled`：

```
# kubectrl get nodes
NAME          STATUS          AGE
k8s-node-1    Ready,SchedulingDisabled  1h
```

对于后续创建的 Pod，系统将不会再向该 Node 进行调度。

也可以不使用配置文件，直接使用 `kubectrl patch` 命令完成：

```
$ kubectrl patch node k8s-node-1 -p '{"spec":{"unschedulable":true}}'
```

需要注意的是，将某个 Node 脱离调度范围时，在其上运行的 Pod 并不会自动停止，管理员需要手动停止在该 Node 上运行的 Pod。

同样，如果需要将某个 Node 重新纳入集群调度范围，则将 `unschedulable` 设置为 `false`，再次执行 `kubectrl replace` 或 `kubectrl patch` 命令就能恢复系统对该 Node 的调度。

在 Kubernetes 当前的版本中，`kubectrl` 的子命令 `cordon` 和 `uncordon` 也用于实现将 Node 进行隔离和恢复调度的操作。

例如，使用 `kubectrl cordon <node_name>` 对某个 Node 进行隔离调度操作：

```
# kubectrl cordon k8s-node-1
node "k8s-node-1" cordoned

# kubectrl get nodes
NAME          STATUS          AGE
k8s-node-1    Ready,SchedulingDisabled  1h
```

使用 `kubectrl uncordon <node_name>` 对某个 Node 进行恢复调度操作：

```
# kubectrl uncordon k8s-node-1
node "k8s-node-1" uncordoned

# kubectrl get nodes
NAME          STATUS    AGE
k8s-node-1    Ready    1h
```

## 2. Node 的扩容

在实际生产系统中会经常遇到服务器容量不足的情况，这时就需要购买新的服务器，然后将应用系统进行水平扩展来完成对系统的扩容。

在 Kubernetes 集群中，一个新 Node 的加入是非常简单的。在新的 Node 节点上安装 Docker、kubelet 和 kube-proxy 服务，然后配置 kubelet 和 kube-proxy 的启动参数，将 Master URL 指定为当前 Kubernetes 集群 Master 的地址，最后启动这些服务。通过 kubelet 默认的自动注册机制，新的 Node 将会自动加入现有的 Kubernetes 集群中，如图 5.1 所示。

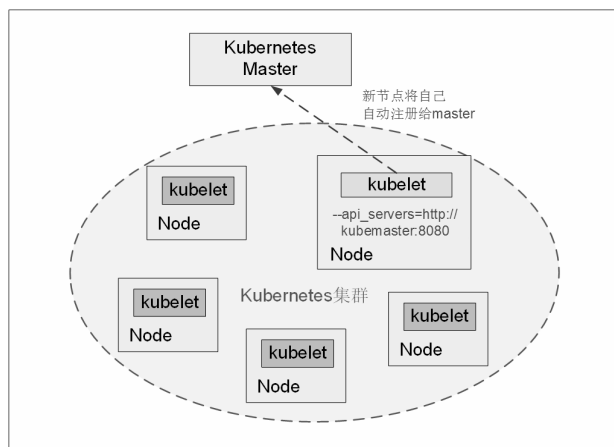


图 5.1 新节点自动注册并加入现有的 Kubernetes 集群中

Kubernetes Master 在接受了新 Node 的注册之后，会自动将其纳入当前集群的调度范围内，在之后创建容器时，就可以向新的 Node 进行调度了。

通过这种机制，Kubernetes 实现了集群中 Node 的扩容。

### 5.1.2 更新资源对象的 Label

Label（标签）作为用户可灵活定义的对象属性，在正在运行的资源对象上，仍然可以随时通过 `kubectl label` 命令对其进行增加、修改、删除等操作。

例如，我们要给已创建的 Pod “redis-master-bobr0” 添加一个标签 `role=backend`：

```
$ kubectl label pod redis-master-bobr0 role=backend
pod "redis-master-bobr0" labeled
```

查看该 Pod 的 Label：

```
$ kubectl get pods -Lrole
```

NAME	READY	STATUS	RESTARTS	AGE	ROLE
redis-master-bobr0	1/1	Running	0	3m	backend

删除一个 Label 时，只需在命令行最后指定 Label 的 key 名并与一个减号相连即可：

```
$ kubectl label pod redis-master-bobr0 role-  
pod "redis-master-bobr0" labeled
```

修改一个 Label 的值时，需要加上--overwrite 参数：

```
$ kubectl label pod redis-master-bobr0 role=master --overwrite  
pod "redis-master-bobr0" labeled
```

### 5.1.3 Namespace：集群环境共享与隔离

在一个组织内部，不同的工作组可以在同一个 Kubernetes 集群中工作，Kubernetes 通过命名空间和 Context 的设置来对不同的工作组进行区分，使得它们既可以共享同一个 Kubernetes 集群的服务，也能够互不干扰，如图 5.2 所示。

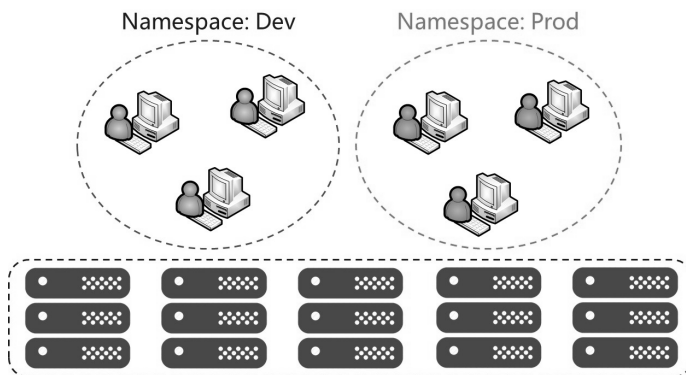


图 5.2 集群环境共享和隔离

假设在我们的组织中有两个工作组：开发组和生产运维组。开发组在 Kubernetes 集群中需要不断创建、修改、删除各种 Pod、RC、Service 等资源对象，以便实现敏捷开发的过程。而生产运维组则需要使用严格的权限设置来确保生产系统中的 Pod、RC、Service 处于正常运行状态且不会被误操作。

#### 1. 创建 namespace

为了在 Kubernetes 集群中实现这两个分组，首先需要创建两个命名空间。

**namespace-development.yaml:**

```
apiVersion: v1
```

```
kind: Namespace
metadata:
  name: development
```

#### **namespace-production.yaml:**

```
apiVersion: v1
kind: Namespace
metadata:
  name: production
```

使用 **kubectl create** 命令完成命名空间的创建：

```
$ kubectl create -f namespace-development.yaml
namespaces/development
```

```
$ kubectl create -f namespace-production.yaml
namespaces/production
```

查看系统中的命名空间：

```
$ kubectl get namespaces
NAME          LABELS              STATUS
default       <none>              Active
development   name=development    Active
production    name=production     Active
```

## **2. 定义 Context（运行环境）**

接下来，需要为这两个工作组分别定义一个 Context，即运行环境。这个运行环境将属于某个特定的命名空间。

通过 **kubectl config set-context** 命令定义 Context，并将 Context 置于之前创建的命名空间中：

```
$ kubectl config set-cluster kubernetes-cluster --server=https://192.168.1.128:8080
$ kubectl config set-context ctx-dev --namespace=development --cluster=kubernetes-cluster --user=dev
$ kubectl config set-context ctx-prod --namespace=production --cluster=kubernetes-cluster --user=prod
```

使用 **kubectl config view** 命令查看已定义的 Context：

```
$ kubectl config view
apiVersion: v1
clusters:
- cluster:
    server: http://192.168.1.128:8080
    name: kubernetes-cluster
```



```

contexts:
- context:
  cluster: kubernetes-cluster
  namespace: development
  name: ctx-dev
- context:
  cluster: kubernetes-cluster
  namespace: production
  name: ctx-prod
current-context: ctx-dev
kind: Config
preferences: {}
users: []

```

注意，通过 `kubectl config` 命令在 `${HOME}/.kube` 目录下生成了一个名为 `config` 的文件，文件内容即以 `kubectl config view` 命令查看到的内容。所以，也可以通过手工编辑该文件的方式来设置 Context。

### 3. 设置工作组在特定 Context 环境中工作

使用 `kubectl config use-context <context_name>` 命令来设置当前的运行环境。

下面的命令将把当前运行环境设置为 “ctx-dev”：

```
$ kubectl config use-context ctx-dev
```

通过这个命令，当前的运行环境即被设置为开发组所需的环境。之后的所有操作都将在名为 “development” 的命名空间中完成。

现在，以 `redis-slave RC` 为例创建两个 Pod：

#### **redis-slave-controller.yaml**

```

apiVersion: v1
kind: ReplicationController
metadata:
  name: redis-slave
  labels:
    name: redis-slave
spec:
  replicas: 2
  selector:
    name: redis-slave
  template:
    metadata:
      labels:
        name: redis-slave
    spec:

```

```
containers:
- name: slave
  image: kubeguide/guestbook-redis-slave
  ports:
  - containerPort: 6379

$ kubectl create -f redis-slave-controller.yaml
replicationcontrollers/redis-slave
```

查看创建好的 Pod:

```
$ kubectl get pods
NAME                READY    STATUS    RESTARTS   AGE
redis-slave-0feq9    1/1     Running   0           6m
redis-slave-6i0g4    1/1     Running   0           6m
```

可以看到容器被正确创建并运行起来了。而且，由于当前的运行环境是 `ctx-dev`，所以不会影响到生产运维组的工作。

让我们切换到生产运维组的运行环境:

```
$ kubectl config use-context ctx-prod
```

查看 RC 和 Pod:

```
$ kubectl get rc
CONTROLLER  CONTAINER(S)  IMAGE(S)  SELECTOR  REPLICAS

$ kubectl get pods
NAME        READY    STATUS    RESTARTS   AGE
```

结果为空，说明看不到开发组创建的 RC 和 Pod。

现在我们为生产运维组也创建两个 `redis-slave` 的 Pod:

```
$ kubectl create -f redis-slave-controller.yaml
replicationcontrollers/redis-slave
```

查看创建好的 Pod:

```
$ kubectl get pods
NAME                READY    STATUS    RESTARTS   AGE
redis-slave-a4m7s    1/1     Running   0           12s
redis-slave-xyrkk    1/1     Running   0           12s
```

可以看到容器被正确创建并运行起来了，并且当前的运行环境是 `ctx-prod`，也不会影响开发组的工作。

至此，我们为两个工作组分别设置了两个运行环境，在设置好当前的运行环境时，各工作组之间的工作将不会相互干扰，并且它们都能够在同一个 Kubernetes 集群中同时工作。

### 5.1.4 Kubernetes 资源管理

本章从计算资源管理（Compute Resources）、资源配置范围管理（LimitRange）、服务质量管理（QoS）及资源配额管理（ResourceQuota）等方面，对 Kubernetes 集群内的资源管理进行详细说明，结合实践操作、常见问题分析和一个完整的示例，对 Kubernetes 集群资源管理相关的运维工作提供指导。

#### 1. 计算资源管理（Compute Resources）

在配置 Pod 时，我们可以为其中的每个容器指定需要使用的计算资源（CPU 和内存）。

计算资源的配置项分为两种：一种是资源请求（Resource Requests，简称 Requests），表示容器希望被分配到的、可完全保证的资源量，Requests 的值会提供给 Kubernetes 调度器（Kubernetes Scheduler）以便于优化基于资源请求的容器调度；另外一种资源限制（Resource Limits，简称 Limits），Limits 是容器最多能使用到的资源量的上限，这个上限值会影响节点上发生资源竞争时的解决策略。

当前版本的 Kubernetes 中，计算资源的资源类型分为两种：CPU 和内存（Memory）。这两种资源类型都有一个基本单位：对于 CPU 而言，基本单位是核心数（Cores）；而内存的基本单位是字节数（Bytes）。CPU 和内存一起构成了目前 Kubernetes 中的计算资源（可简称资源）。

计算资源是可计量的，能被申请、分配和使用的基础资源，这使之区别于 API 资源（API Resources，例如 Pod 和 services 等）。

#### 1) Pod 和容器的 Requests 和 Limits

Pod 中的每个容器都可以配置以下 4 个参数。

- ◎ `spec.container[].resources.requests.cpu`。
- ◎ `spec.container[].resources.limits.cpu`。
- ◎ `spec.container[].resources.requests.memory`。
- ◎ `spec.container[].resources.limits.memory`。

这四个参数分别对应容器的 CPU 和内存的 Requests 和 Limits，它们具有以下特点。

- ◎ Requests 和 Limits 都是可选的。在某些集群中如果在 Pod 创建或者更新时，没设置资源限制或者资源请求值，那么可能会使用系统提供一个默认值，这个默认值取决于集群的配置。
- ◎ 如果 Request 没有配置，那么默认会被设置为等于 Limits。
- ◎ 而任何情况下 Limits 都应该设置为大于或者等于 Requests。

以 CPU 为例，图 5.3 显示了未设置 CPU Limits 和设置 CPU Limits 的 CPU 使用率的区别。

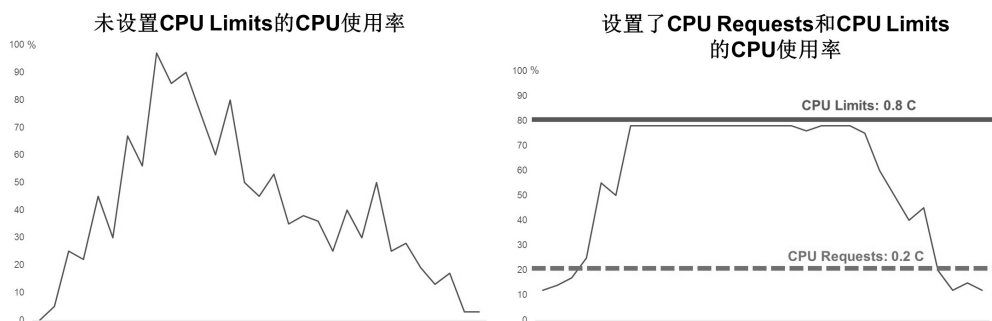


图 5.3 未设置和设置了 CPU Limits 的 CPU 使用率的区别

尽管 Requests 和 Limits 只能设置到容器上，但是设置 Pod 级别的 Requests 和 Limits 能很大程度上提高我们对 Pod 管理的便利性和灵活性，因此 Kubernetes 中提供对 Pod 级别的 Requests 和 Limits 配置。对于 CPU 和内存而言，Pod 的 Requests 或 Limits 是指该 Pod 中所有容器的 Requests 或 Limits 的总和（Pod 中没设置 Request 或 Limits 的容器，该项的值被当作 0 或者按照集群配置的默认值来计算）。下面对 CPU 和内存这两种计算资源各自的特点进行说明。

### （1）CPU

CPU 的 Requests 和 Limits 是通过 CPU 数（cpus）来度量的。CPU 资源值支持最多三位小数：如果一个容器的 `spec.container[].resources.requests.cpu` 设置为 0.5，那么它会获得半个 CPU；同理如果设置为 1，就会获得 1 个 CPU。0.1CPU 等价于 100m CPU（100 millicpu），而在 Kubernetes API 中自动将这种小数 0.1 转化为 100m，因此 CPU 的小数最多支持三位数字，而 Kubernetes 官方也更推荐直接使用形如 100m 的 millicpu 作为计量单位。

CPU 资源值是绝对值，而不是相对值：比如 0.1CPU 不管是在单核或者多核机器上都是一样的，都严格等于 0.1 CPU core。

### （2）内存（Memory）

内存的 Requests 和 Limits 计量单位是字节数（Bytes）。内存值用使用整数或者定点整数加上国际单位制（International System of Units）来表示。国际单位制包括十进制的 E、P、T、G、M、K、m，或二进制的 Ei、Pi、Ti、Gi、Mi、Ki。比如：KiB 与 MiB 是二进制表示的字节单位，而常见的 KB 与 MB 则是十进制表示的字节单位。两种方式的区别举例说明如下：

1 KB（kilobyte）= 1000 bytes = 8000 bits

1 KiB（kibibyte）=  $2^{10}$  bytes = 1024 bytes = 8192 bits

因此，下面几种内存配置的意思是一样的：128974848、129e6、129M、123Mi

Kubernetes 的计算资源单位是大小写敏感的，因为 m 可以表示千分之一单位（milli unit），而 M 可以表示十进制的 1000，两者的含义不同；同理可知，小写的 k 不是一个合法的资源单位。

以一个 Pod 中的资源配置为例：

```
apiVersion: v1
kind: Pod
metadata:
  name: frontend
spec:
  containers:
  - name: db
    image: mysql
    resources:
      requests:
        memory: "64Mi"
        cpu: "250m"
      limits:
        memory: "128Mi"
        cpu: "500m"
  - name: wp
    image: wordpress
    resources:
      requests:
        memory: "64Mi"
        cpu: "250m"
      limits:
        memory: "128Mi"
        cpu: "500m"
```

该 Pod 包含两个容器，每个容器配置的 Requests 都是 0.25 CPU 和 64MiB ( $2^{26}$  bytes) 内存，而配置的 Limits 都是 0.5 CPU 和 128MiB ( $2^{27}$  bytes) 内存。

这个 Pod 的 Requests 和 Limits 等于 Pod 中所有容器对应配置的总和，所以 Pod 的 Requests 是 0.5 CPU 和 128MiB ( $2^{27}$  bytes) 内存，Limits 是 1 CPU 和 256MiB ( $2^{28}$  bytes) 内存。

## 2) 基于 Requests 和 Limits 的 Pod 调度机制

当一个 Pod 创建成功时，Kubernetes 调度器（Scheduler）为该 Pod 选择一个节点（Node）来执行。对于每种计算资源（CPU 和内存）而言，每个节点都有一个能用于运行 Pod 的最大容量值。调度器在调度时，首先要确保调度后该节点上所有 Pod 的 CPU 和内存的 Requests 总和不能超过该节点能提供给 Pod 使用的 CPU 和内存的最大容量值。

例如某个节点上 CPU 资源充足，而内存为 4GB，其中 3GB 可以运行 Pod，某 Pod 的内存 Requests 为 1GB、Limits 为 2GB，那么这个节点上最多可运行 3 个这种 Pod。

这里需要注意的是：可能某些节点上的实际资源使用量非常低，但是如果该节点上已运行 Pod 配置的 Requests 值的总和已经非常高，再加上需要调度的 Pod 的 Requests 值会直接超过该节点提供给 Pod 的资源容量上限，Kubernetes 仍然不会将 Pod 调度到这个节点上。这是因为如果 Kubernetes 将 Pod 调度到该节点上，那么如果后面该节点上运行的 Pod 面临服务峰值等情况，可能会导致 Pod 资源短缺的情况发生。

接着上面的例子，假设该节点已经启动 3 个 Pod 实例，而这 3 个 Pod 的实际内存使用都不足 500MB，那么理论上该节点的可用内存应该大于 1.5GB，但是由于该节点的 Pod Requests 总和已经等于节点的可用内存上限，因此 Kubernetes 不会再将任何 Pod 实例调度到该节点上执行。

### 3) Requests 和 Limits 资源配置机制

当 kubelet 启动 Pod 的一个容器时，它会将容器的 Requests 和 Limits 值转化为相应的容器启动参数传递给容器执行器（Docker 或者是 rkt）。

如果容器的执行环境是 Docker，那么容器的 4 个参数是这样传递给 Docker 的。

#### (1) spec.container[].resources.requests.cpu

这个参数会转化为 core 数（比如配置的 100m 会转化为 0.1），然后乘以 1024，再将这个结果作为 --cpu-shares 参数的值传递给 docker run 命令。在 docker run 命令中，--cpu-share 参数是一个相对权重值（Relative Weight），这个相对权重值会决定 Docker 在资源竞争时分配给容器的资源比例。举例说明 --cpu-shares 参数在 Docker 中的含义：比如两个容器的 CPU Requests 分别设置为 1 和 2，那么容器在 docker run 启动时对应的 --cpu-shares 参数值分别为 1024 和 2048，在主机 CPU 资源产生竞争时，Docker 会尝试按照 1：2 的配比将 CPU 资源分配给这两个容器使用。

这里需要区分清楚的是：这个参数对于 Kubernetes 而言是绝对值，主要用于 Kubernetes 调度和管理的依据（参见下文 QoS 章节）；同时这个参数值会设置为 --cpu-shares 参数传递给 Docker，--cpu-shares 参数对于 Docker 而言又是相对值，主要用于资源分配比例。这两种用途的作用范围不同，所以并不会发生冲突。

#### (2) spec.container[].resources.limits.cpu

这个参数会转化为 millicore 数（比如配置的 1 会转化为 1000，而配置的 100m 转化为 100），将此值乘以 100000，再除以 1000，然后将结果值作为 --cpu-quota 参数的值传递给 docker run 命令。docker run 命令中另外一个参数 --cpu-period 默认设置为 100000，表示 Docker 重新计量和分配 CPU 的使用时间间隔为 100000μs（100ms）。

Docker 的 --cpu-quota 参数和 --cpu-period 参数一起配合完成对容器 CPU 的使用限制：比如

Kubernetes 中配置容器的 CPU Limits 为 0.1，那么计算后--cpu-quota 为 10000，而--cpu-period 为 100000，这意味着 Docker 在 100ms 内最多给该容器分配  $10\text{ms} \times \text{core}$  的计算资源用量， $10/100=0.1$  core 的结果与 Kubernetes 配置的意义是一致的。

注意：如果 kubelet 启动参数--cpu-cfs-quota 设置为 true，那么 kubelet 会强制要求所有 Pod 都必须配置 CPU Limits（如果 Pod 没配置，而集群提供了默认配置也可以）。而从 Kubernetes 1.2 版本开始，这个--cpu-cfs-quota 启动参数的默认值就是 true。

### （3）spec.container[].resources.requests.memory

这个参数值只提供给 Kubernetes 调度器（Kubernetes Scheduler）作为调度和管理的依据，不会作为任何参数传递给 Docker。

### （4）spec.container[].resources.limits.memory

这个参数值会转化为单位为 Bytes 的整数，数值会作为--memory 参数传递给 docker run 命令。

如果一个容器在运行过程中使用了超出了其内存 Limits 配置的内存限制值，那么它可能会被“杀掉”，如果这个容器是一个可重启的容器，那么之后它会被 kubelet 重新启动起来。因此容器的 Limits 配置需要进行准确的测试和评估。

与内存 Limits 不同的是 CPU 在容器技术中属于可压缩资源，因此对于 CPU 的 Limits 配置一般不会引发因偶然超标使用而导致容器被系统“杀掉”的情况。

## 4) 计算资源使用情况监控

Pod 的资源用量会作为 Pod 的状态信息一同上报给 Master。如果集群中配置了 Heapster 来监控集群的性能数据，那么还可以从 Heapster 中查看 Pod 的资源用量信息。

## 5) 计算资源相关常见问题分析

（1）Pod 状态为 pending，错误信息为 FailedScheduling。如果 Kubernetes 调度器（Kubernetes Scheduler）在集群中找不到合适的节点来运行 Pod，那么这个 Pod 会一直处于未调度状态，直到调度器找到合适的节点为止。每次调度器尝试调度失败，Kubernetes 都会产生一个事件（event），我们可以通过下面这种方式来查看事件的信息：

```
$ kubectl describe pod frontend | grep -A 3 Events
Events:
  FirstSeen    LastSeen    Count  From              Subobject    PathReason    Message
  36s          5s          6      {scheduler }      FailedScheduling Failed for reason PodExceedsFreeCPU and possibly others
```

在上面这个例子中，名为 frontend 的 Pod 由于节点的 CPU 资源不足而调度失败（Pod ExceedsFreeCPU），同样，如果内存不足也可能导致调度失败（PodExceedsFreeMemory）。

如果一个或者多个 Pod 调度失败且有这类错误，那么我们可以尝试以下几种解决方法。

- ◎ 添加更多的节点到集群中。
- ◎ 停止一些不必要的运行中的 Pod，释放资源。
- ◎ 检查 Pod 的配置，错误的配置可能导致该 Pod 永远都无法被调度执行。比如如果整个集群中所有节点都只有 1 CPU，而 Pod 配置的 CPU Requests 为 2，那么该 Pod 就不会被调度执行。

我们可以使用 `kubectl describe nodes` 命令来查看集群中节点的计算资源容量和已使用量：

```
$ kubectl describe nodes k8s-node-1
Name:          k8s-node-1
...
Capacity:
  cpu:          1
  memory:       464Mi
  pods:         40
Allocated resources (total requests):
  cpu:          910m
  memory:       2370Mi
  pods:         4
...
Pods:          (4 in total)
  Namespace    Name                                CPU(millicPU)
Memory(bytes)
  frontend     webserver-ffj8j                    500 (50% of total)
  2097152000 (50% of total)
  kube-system  fluentd-cloud-logging-k8s-node-1   100 (10% of total)
  209715200 (5% of total)
  kube-system  kube-dns-v8-qopgw                  310 (31% of total)
178257920 (4% of total)
TotalResourceLimits:
  CPU(millicPU):    910 (91% of total)
  Memory(bytes):    2485125120 (59% of total)
...
```

超过可用资源容量上限（Capacity）和已分配资源量（Allocated resources）差额的 Pod 无法运行在该 Node 上。这个例子中，如果一个 Pod 的 Requests 超过 90 millicpus 或者超过 1341MiB 内存，那么就无法运行在这个节点上。

在后面的资源配额（Resource Quota）章节中，我们还可以配置针对一组 Pod 的 Requests 和 Limits 总量的限制，这种限制可以作用于命名空间，通过这种方式我们可以防止一个命名空间下的用户将所有资源全部据为己有。

（2）容器被强行终止（Terminated）。如果容器使用的资源超过了它配置的 Limits，那么该



容器可能会被强制终止。我们可以通过 `kubectl describe pod` 命令来确认容器是否因为这个原因被终止：

```
$ kubectl describe pod simmemleak-hra99
Name:                simmemleak-hra99
Namespace:           default
Image(s):            saadali/simmemleak
Node:                192.168.18.3
Labels:              name=simmemleak
Status:              Running
Reason:
Message:
IP:                  172.17.1.3
Replication Controllers:  simmemleak (1/1 replicas created)
Containers:
  simmemleak:
    Image: saadali/simmemleak
    Limits:
      cpu:                100m
      memory:             50Mi
    State: Running
      Started:            Tue, 07 Jul 2015 12:54:41 -0700
    Last Termination State: Terminated
      Exit Code:          1
      Started:            Fri, 07 Jul 2015 12:54:30 -0700
      Finished:           Fri, 07 Jul 2015 12:54:33 -0700
    Ready:                False
    Restart Count:        5
Conditions:
  Type      Status
  Ready     False
Events:
  FirstSeen          LastSeen          Count  From
SubobjectPath      Reason      Message
  Tue, 07 Jul 2015 12:53:51 -0700    Tue, 07 Jul 2015 12:53:51 -0700    1
{scheduler }
Successfully assigned simmemleak-hra99 to kubernetes-node-tf0f
  Tue, 07 Jul 2015 12:53:51 -0700    Tue, 07 Jul 2015 12:53:51 -0700    1    {kubelet
kubernetes-node-tf0f}    implicitly required container POD    pulled    Pod
container image "gcr.io/google_containers/pause:0.8.0" already present on machine
  Tue, 07 Jul 2015 12:53:51 -0700    Tue, 07 Jul 2015 12:53:51 -0700    1    {kubelet
kubernetes-node-tf0f}    implicitly required container POD    created    Created
with docker id 6a41280f516d
  Tue, 07 Jul 2015 12:53:51 -0700    Tue, 07 Jul 2015 12:53:51 -0700    1    {kubelet
kubernetes-node-tf0f}    implicitly required container POD    started    Started
with docker id 6a41280f516d
  Tue, 07 Jul 2015 12:53:51 -0700    Tue, 07 Jul 2015 12:53:51 -0700    1    {kubelet
```

```
kubernetes-node-tf0f}    spec.containers{simmemleak}          created    Created
with docker id 87348f12526a
```

Restart Count: 5 说明这个名为 `simmemleak` 的容器被强制终止并重启了 5 次。

我们可以在使用 `kubectl get pod` 命令时添加 `-o go-template=...` 格式参数来读取已终止容器之前的状态信息：

```
$ kubectl get pod -o
go-template='{{range.status.containerStatuses}}{{"Container Name:
"}}{{.name}}{{"\r\nLastState: "}}{{.lastState}}{{end}}' simmemleak-60xbc
Container Name: simmemleak
LastState: map[terminated:map[exitCode:137 reason:OOM Killed
startedAt:2015-07-07T20:58:43Z finishedAt:2015-07-07T20:58:43Z
containerID:docker://0e4095bba1feccdfe7ef9fb6ebffe972b4b14285d5acdec6f0d3ae8a22fad8b2]]
```

这里我们可以看到这个容器因为 `reason:OOM Killed` 而被强制终止，说明这个容器的内存超过了限制（Out of Memory）。

## 6) 计算资源管理的演进

当前版本的 Kubernetes 中的 Requests 和 Limits 都是作用于容器级别的，未来 Kubernetes 计划增加对直接作用于 Pod 级别的资源配置的支持，这种资源配置是能被 Pod 内的所有容器共享的，包括 `emptyDir` 这种 Pod 级别的 Volume。

从资源的种类来看，目前 Kubernetes 只能支持 CPU 和内存两种计算资源类型，在后续的版本中，Kubernetes 计划支持更多的资源类型，包括节点磁盘空间资源，还将支持自定义的资源类型。

## 2. 资源的配置范围管理（LimitRange）

默认情况下，Kubernetes 的 Pod 会以无限制的 CPU 和内存运行。这也就意味着 Kubernetes 系统中任何的 Pod 都可以使用其所在节点上的所有可用的 CPU 和内存。通过配置 Pod 的计算资源 Requests 和 Limits，我们可以限制 Pod 的资源使用，但对于 Kubernetes 集群管理员而言，配置每一个 Pod 的 Requests 和 Limits 是烦琐且限制性过强的。更多时，我们需要的是对集群内 Request 和 Limits 的配置做一个全局的统一的限制。常见的配置场景如下。

- ◎ 集群中的每个节点有 2GB 内存，集群管理员不希望任何 Pod 申请超过 2GB 的内存：因为整个集群中没有任何节点能满足超过 2GB 内存的请求。如果某个 Pod 的内存配置超过 2GB，那么该 Pod 将永远都无法被调度到任何节点上执行。为了防止这种情况的发生，集群管理员希望能在系统管理功能中设置禁止 Pod 申请超过 2GB 内存。
- ◎ 集群由同一个组织中的两个团队共享，各自分别用来运行生产环境和开发环境。生产环境最多可以使用 8GB 内存，而开发环境最多可以使用 512MB 内存。集群管理员希

望通过为这两个环境创建不同的命名空间（namespace）并为每个命名空间设置不同的限制来满足这个需求。

- ◎ 用户创建 Pod 时使用的资源可能会刚好比整个机器资源的上限稍小一点，而恰好剩下的资源大小非常尴尬：不足以运行其他任务但整个集群加起来又非常浪费。因此，集群管理员希望设置每个 Pod 必须至少使用集群平均资源值（CPU 和内存）的 20%，这样集群能够提供更好的资源一致性的调度，从而减少了资源浪费。

针对这些需求，Kubernetes 提供了 LimitRange 机制对 Pod 和容器的 Requests 和 Limits 配置进一步做出限制。在下面的示例中，将说明如何将 LimitsRange 应用到一个 Kubernetes 的命名空间（namespace）中，然后说明 LimitRange 的几种限制方式，比如最大及最小范围、Requests 和 Limits 的默认值、Limits 与 Requests 最大比例上限等。

下面通过 LimitRange 的设置和应用对其进行说明。

### 1) 创建一个 namespace

创建一个名为 limit-example 的 namespace:

```
$ kubectl create namespace limit-example
namespace "limit-example" created
```

### 2) 为 namespace 设置 LimitRange

为 namespace “limit-example” 创建一个简单的 LimitRange。创建 limits.yaml 配置文件，内容如下：

```
apiVersion: v1
kind: LimitRange
metadata:
  name: mylimits
spec:
  limits:
  - max:
      cpu: "4"
      memory: 2Gi
    min:
      cpu: 200m
      memory: 6Mi
    maxLimitRequestRatio:
      cpu: 3
      memory: 2
    type: Pod
  - default:
      cpu: 300m
      memory: 200Mi
    defaultRequest:
```

```
cpu: 200m
memory: 100Mi
max:
  cpu: "2"
  memory: 1Gi
min:
  cpu: 100m
  memory: 3Mi
maxLimitRequestRatio:
  cpu: 5
  memory: 4
type: Container
```

创建该 LimitRange:

```
$ kubectl create -f limits.yaml --namespace=limit-example
limitrange "mylimits" created
```

查看 namespace limit-example 中的 LimitRange:

```
$ kubectl describe limits mylimits --namespace=limit-example
```

```
Name:      mylimits
Namespace: limit-example
```

Type	Resource	Min	Max	Default Request	Default Limit
Max Limit/Request Ratio					
---	-----	---	---	-----	-----
Pod	cpu	200m	4	-	3
Pod	memory	6Mi	2Gi	-	2
Container	cpu	100m	2	200m	5
Container	memory	3Mi	1Gi	100Mi	4

下面解释一下 LimitRange 中各项配置的意义和特点。

(1) 不论是 CPU 还是内存，在 LimitRange 中，Pod 和 Container 都可以设置 Min、Max 和 Max Limit/Requests Ratio 这三种参数。Container 还可以设置 Default Request 和 Default Limit 这两种参数，而 Pod 不能设置 Default Request 和 Default Limit。

(2) 对 Pod 和 Container 的五种参数的解释如下。

- Container 的 Min(上面的 100m 和 3Mi)是 Pod 中所有容器的 Requests 值的下限; Container 的 Max (上面的 2 和 1Gi) 是 Pod 中所有容器的 Limits 值的上限; Container 的 Default Request(上面的 200m 和 100Mi)是 Pod 中所有未指定 Requests 值的容器的默认 Requests 值; Container 的 Default Limit (上面的 300m 和 200Mi) 是 Pod 中所有未指定 Limits 值的容器的默认 Limits 值。对于同一资源类型，这 4 个参数必须满足以下关系: Min ≤ Default Request ≤ Default Limit ≤ Max。
- Pod 的 Min (上面的 200m 和 6Mi) 是 Pod 中所有容器的 Requests 值的总和的下限; Pod

的 Max（上面的 4 和 2Gi）是 Pod 中所有容器的 Limits 值的总和的上限。当容器未指定 Requests 值或者 Limits 值时，将使用 Container 的 Default Request 值或者 Default Limit 值。

- ◎ Container 的 Max Limit/Requests Ratio（上面的 5 和 4）限制了 Pod 中所有容器的 Limits 值与 Requests 值的比例上限；而 Pod 的 Max Limit/Requests Ratio（上面的 3 和 2）限制了 Pod 中所有容器的 Limits 值总和与 Requests 值总和的比例上限。

（3）如果设置了 Container 的 Max，那么对于该类资源而言，整个集群中的所有容器都必须设置 Limits，否则将无法成功创建。Pod 内的容器未配置 Limits 时，将使用 Default Limit 的值（本例中的 300m CPU 和 200Mi 内存），而如果 Default 也未配置则无法成功创建。

（4）如果设置了 Container 的 Min，那么对于该类资源而言，整个集群中的所有容器都必须设置 Requests。如果创建 Pod 的容器时未配置该类资源的 Requests，那么创建过程会报验证错误。Pod 里容器的 Requests 在未配置时，可以使用默认值 defaultRequest（本例中的 200m CPU 和 100Mi 内存）；如果未配置而又没有 defaultRequest，那么会默认等于该容器的 Limits；如果此时 Limits 也未定义，那么就会报错。

（5）对于任意一个 Pod 而言，该 Pod 中所有容器的 Requests 总和必须大于或等于 6Mi，而且所有容器的 Limits 总和必须小于或等于 1Gi；同样，所有容器的 CPU Requests 总和必须大于或等于 200m，而且所有容器的 CPU Limits 总和必须小于或等于 2。

（6）Pod 里任何容器的 Limits 与 Requests 的比例不能超过 Container 的 Max Limit/Requests Ratio；Pod 里所有容器的 Limits 总和与 Requests 的总和的比例不能超过 Pod 的 Max Limit/Requests Ratio。

### 3) 创建 Pod 时触发 LimitRange 限制

最后，让我们看看 LimitRange 生效时对容器的资源限制效果。

命名空间中的限制（LimitRange）只会在 Pod 创建或者更新时执行检查。如果手动修改限制（LimitRange）为一个新的值，那么这个新的值不会去检查或限制之前已经在该命名空间中创建好的 Pod。

如果用户创建 Pod 时，配置的资源值（CPU 或者内存）超过了 LimitRange 的限制，那么该创建过程会报错，在错误信息中会说明详细的错误原因。

下面通过创建一个单容器 Pod 来展示默认限制是如何配置到 Pod 上的：

```
$ kubectl run nginx --image=nginx --replicas=1 --namespace=limit-example
deployment "nginx" created
```

查看已创建的 Pod：

```
$ kubectl get pods --namespace=limit-example
NAME                                READY    STATUS    RESTARTS    AGE
```

```
nginx-2040093540-s8vzu 1/1 Running 0 11s
```

查看该 Pod 的 resources 相关信息：

```
$ kubectl get pods nginx-2040093540-s8vzu --namespace=limit-example -o yaml |
grep resources -C 8
  resourceVersion: "57"
  selfLink: /api/v1/namespaces/limit-example/pods/nginx-2040093540-ivimu
  uid: 67b20741-f53b-11e5-b066-64510658e388
spec:
  containers:
  - image: nginx
    imagePullPolicy: Always
    name: nginx
    resources:
      limits:
        cpu: 300m
        memory: 200Mi
      requests:
        cpu: 200m
        memory: 100Mi
    terminationMessagePath: /dev/termination-log
    volumeMounts:
```

由于该 Pod 未配置资源 Requests 和 Limits，所以使用了 namespace limit-example 中的默认 CPU 和内存定义的 Requests 和 Limits 值。

下面创建一个超出资源限制的 Pod（使用 3 CPU）：

```
invalid-pod.yaml:
apiVersion: v1
kind: Pod
metadata:
  name: invalid-pod
spec:
  containers:
  - name: kubernetes-serve-hostname
    image: gcr.io/google_containers/serve_hostname
    resources:
      limits:
        cpu: "3"
        memory: 100Mi
```

创建该 Pod，可以看到系统报错了，并且提供了错误原因为超过了限制。

```
$ kubectl create -f invalid-pod.yaml --namespace=limit-example
Error from server: error when creating "invalid-pod.yaml": Pod "invalid-pod" is
forbidden: [Maximum cpu usage per Pod is 2, but limit is 3., Maximum cpu usage per
Container is 2, but limit is 3.]
```

接下来的例子展示了 LimitRange 对 maxLimitRequestRatio 的限制：

```
limit-test-nginx.yaml:
apiVersion: v1
kind: Pod
metadata:
  name: limit-test-nginx
  labels:
    name: limit-test-nginx
spec:
  containers:
  - name: limit-test-nginx
    image: nginx
    resources:
      limits:
        cpu: "1"
        memory: 512Mi
      requests:
        cpu: "0.8"
        memory: 250Mi
```

由于 limit-test-nginx 这个 Pod 的全部内存 Limits 总和与 Requests 总和的比例为 512 : 250, 大于 LimitRange 中定义的 Pod 的内存 maxLimitRequestRatio 值 2, 因此创建会失败:

```
$ kubectl create -f limit-test-nginx.yaml --namespace=limit-example
Error from server: error when creating "limit-test-nginx.yaml": pods
"limit-test-nginx" is forbidden: [memory max limit to request ratio per Pod is 2,
but provided ratio is 2.048000.]
```

下面的例子为满足 LimitRange 限制的 Pod:

```
valid-pod.yaml:
apiVersion: v1
kind: Pod
metadata:
  name: valid-pod
  labels:
    name: valid-pod
spec:
  containers:
  - name: kubernetes-serve-hostname
    image: gcr.io/google_containers/serve_hostname
    resources:
      limits:
        cpu: "1"
        memory: 512Mi
```

创建 Pod 将会成功:

```
$ kubectl create -f valid-pod.yaml --namespace=limit-example
```

```
pod "valid-pod" created
```

查看该 Pod 的资源信息：

```
$ kubectl get pods valid-pod --namespace=limit-example -o yaml | grep -C 6
resources
  uid: 3b1bfd7a-f53c-11e5-b066-64510658e388
spec:
  containers:
  - image: gcr.io/google_containers/serve_hostname
    imagePullPolicy: Always
    name: kubernetes-serve-hostname
    resources:
      limits:
        cpu: "1"
        memory: 512Mi
      requests:
        cpu: "1"
        memory: 512Mi
```

可以看到该 Pod 配置了明确的 Limits 和 Requests，因此该 Pod 不会使用 namespace limit-example 中定义的 default 和 defaultRequest。

需要注意的是，CPU Limits 强制配置这个选项在 Kubernetes 集群中默认是开启的；除非集群管理员在部署 kubelet 时，通过设置参数--cpu-cfs-quota=false 来关闭该限制：

```
$ kubelet --help
Usage of kubelet
....
--cpu-cfs-quota[=true]: Enable CPU CFS quota enforcement for containers that
specify CPU limits
$ kubelet --cpu-cfs-quota=false ...
```

如果集群管理员希望对整个集群中容器或者 Pod 配置的 Requests 和 Limits 做限制，那么可以通过配置 Kubernetes 的命名空间（namespace）上的 LimitRange（资源限制区间）来达到该目的。在 Kubernetes 集群中，如果 Pod 没有显式定义 Limits 和 Requests，那么 Kubernetes 系统会将该 Pod 所在的命名空间中定义的 LimitRange 的 default 和 defaultRequests 配置到该 Pod 上。

### 3. 资源的服务质量管理（Resource QoS）

本节对 Kubernetes 如何根据 Pod 的 Requests 和 Limits 配置来实现针对 Pod 的不同级别的资源服务质量控制（QoS）进行说明。

在 Kubernetes 的资源 QoS 体系中，需要保证高可靠性的 Pod 可以申请可靠资源，而一些不需要高可靠性的 Pod 可以申请可靠性较低或者不可靠的资源。在计算资源一节中，我们讲到了容器的资源配置分为 Requests 和 Limits，其中 Requests 是 Kubernetes 调度时能为容器提供的完



全可保障的资源量（最低保障），而 Limits 是系统允许容器运行时可能使用到的资源量的上限（最高上限）。Pod 级别的资源配置是通过计算 Pod 内所有容器的资源配置的总和得出来的。

Kubernetes 中 Pod 的 Requests 和 Limits 资源配置有如下特点：如果 Pod 配置的 Requests 值等于 Limits 值，那么该 Pod 可以获得的资源是完全可靠的；而如果 Pod 的 Requests 值小于 Limits 值，那么该 Pod 获得的资源可分成两部分：一部分是完全可靠的资源，资源量大小等于 Requests 值；另外一部分是不可靠的资源，这部分资源最大等于 Limits 与 Requests 的差额值，这份不可靠的资源能够申请到多少，则取决于当时主机上容器可用资源的余量。

通过这种机制，Kubernetes 可以实现节点资源的超售（Over Subscription），比如在 CPU 完全充足的情况下，某机器共有 32GiB 内存可提供给容器使用，容器配置为 Requests 值 1GiB，Limits 值为 2GiB，那么该机器上最多可以同时运行 32 个容器，每个容器最多可使用 2GiB 内存，如果这些容器的内存使用峰值错开，那么所有容器也可以一直正常运行。

超售机制能有效地提高资源的利用率，同时不会影响容器申请的完全可靠资源的可靠性。

### 1) Requests 和 Limits 对不同计算资源类型的限制机制

根据计算资源章节的内容我们知道，容器的资源配置满足以下两个条件。

- ◎ Requests ≤ 节点可用资源。
- ◎ Requests ≤ Limits。

Kubernetes 根据 Pod 配置的 Requests 值来调度 Pod，Pod 在成功调度之后会得到 Requests 值定义的资源来运行；而如果 Pod 所在机器上的资源有空余，则 Pod 可以申请更多的资源，最多不能超过 Limits 的值。我们下面看一下 Requests 和 Limits 针对不同计算资源类型的限制机制的差异。这种差异主要取决于计算资源类型是可压缩资源还是不可压缩资源。

#### (1) 可压缩资源

- ◎ Kubernetes 目前支持的可压缩资源是 CPU。
- ◎ Pod 可以得到 Pod 的 Requests 配置的 CPU 使用量，而是否能使用超过 Requests 值的部分取决于系统的负载和调度。不过由于目前 Kubernetes 和 Docker 的 CPU 隔离机制都是在容器级别隔离的，所以 Pod 级别的资源配置并不能完全得到保障；Pod 级别的 cgroups 正在紧锣密鼓地开发中，如果将来引入，就可以确保 Pod 级别的资源配置准确运行。
- ◎ 空闲 CPU 资源按照容器 Requests 值的比例分配。举例说明：容器 A 的 CPU 配置为 Requests 1 Limits 10，容器 B 的 CPU 配置为 request 2 Limits 8，A 和 B 同时运行在一个节点上，初始状态下容器的可用 CPU 为 3cores，那么 A 和 B 恰好得到它们的 Requests 中定义的 CPU 用量，即 1CPU 和 2CPU。如果 A 和 B 都需要更多的 CPU 资源，而恰

好此时系统的其他任务释放出 1.5CPU，那么这 1.5CPU 将按照 A 和 B 的 Requests 值的比例 1：2 分配给 A 和 B，即最终 A 可使用 1.5CPU，B 可使用 3CPU。

- ◎ 如果 Pod 使用了超过 Limits 10 中配置的 CPU 用量，那么 cgroups 会对 Pod 中的容器的 CPU 使用进行限流（throttled）；如果 Pod 没有配置 Limits 10，那么 Pod 会尝试抢占所有空闲的 CPU 资源（Kubernetes 从 v1.2 版本开始默认开启--cpu-cfs-quota，因此默认情况下必须配置 Limits）。

## （2）不可压缩资源

- ◎ Kubernetes 目前支持的可压缩资源是内存。
- ◎ Pod 可以得到 Requests 中配置的内存。如果 Pod 使用的内存量小于它的 Requests 的配置，那么这个 Pod 可以正常运行（除非出现操作系统级别的内存不足等严重问题）；如果 Pod 使用的内存量超过了它的 Requests 的配置，那么这个 Pod 有可能被 Kubernetes “杀掉”：比如 Pod A 使用了超过 Requests 而不到 Limits 的内存量，此时同一机器上另外一个 Pod B 之前只使用了远少于自己的 Requests 值的内存，而此时程序压力增大，Pod B 向系统申请的总量不超过自己的 Requests 值的内存，那么 Kubernetes 可能会直接杀掉 Pod A；另外一种情况是 Pod A 使用了超过 Requests 而不到 Limits 的内存量，此时 Kubernetes 将一个新的 Pod 调度到这台机器上，新的 Pod 需要使用内存，而只有 Pod A 使用了超过了自己的 Requests 值的内存，那么 Kubernetes 也可能会杀掉 Pod A 来释放内存资源。
- ◎ 如果 Pod 使用的内存量超过了它的 Limits 设置，那么操作系统内核会杀掉 Pod 所有容器的所有进程中使用内存最多的一个，直到内存不超过 Limits 为止。

## 2）对调度策略的影响

- ◎ Kubernetes 的 kubelet 通过计算 Pod 中所有容器的 Requests 的总和来决定对 Pod 的调度。
- ◎ 不管是 CPU 还是内存，Kubernetes 调度器和 kubelet 都会确保节点上所有 Pod 的 Requests 的总和不会超过该节点上可分配给容器使用的资源容量上限。

## 3）服务质量等级（QoS Classes）

在一个超用（Over Committed，容器 Limits 总和大于系统容量上限）系统中，由于容器负载的波动可能导致操作系统的资源不足，最终可能会导致部分容器被“杀掉”。在这种情况下，我们当然会希望优先“杀掉”那些不太重要的容器，那么如何衡量重要程度呢？Kubernetes 将容器划分成 3 个 QoS 等级：Guaranteed（完全可靠的）、Burstable（弹性波动、较可靠的）和 Best-Effort（尽力而为、不太可靠的），这三种优先级依次递减，如图 5.4 所示。

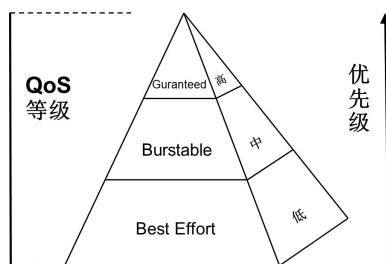


图 5.4 QoS 等级和优先级的关系

从理论上来说，QoS 级别应该作为一个单独的参数来提供 API，并由用户对 Pod 进行配置，这种配置应该与 Requests 和 Limits 无关。但在当前版本的 Kubernetes 的设计中，为了简化模式及避免引入太多的复杂性，QoS 级别直接由 Requests 和 Limits 来定义。在 Kubernetes 中容器的 QoS 级别等于容器所在 Pod 的 QoS 级别，而 Kubernetes 的资源配置定义了 Pod 的三种 QoS 级别，如下所述。

### 1) Guaranteed（完全可靠的）

如果 Pod 中的所有容器对所有资源类型都定义了 Limits 和 Requests，并且所有容器的 Limits 值都和 Requests 值全部相等（且都不为 0），那么该 Pod 的 QoS 级别就是 Guaranteed。注意：在这种情况下，容器可以不定义 Requests，因为 Requests 值在未定义时默认等于 Limits。

下面这两个例子中定义的 Pod QoS 级别就是 Guaranteed:

```
containers:
  name: foo
    resources:
      limits:
        cpu: 10m
        memory: 1Gi
  name: bar
    resources:
      limits:
        cpu: 100m
        memory: 100Mi
```

在上面的例子中未定义 Requests 值，所以其默认等于 Limits 值。而下面这个例子中定义的 Requests 和 Limits 的值完全相同:

```
containers:
  name: foo
    resources:
      limits:
        cpu: 10m
        memory: 1Gi
```

```
      requests:
        cpu: 10m
        memory: 1Gi
name: bar
  resources:
    limits:
      cpu: 100m
      memory: 100Mi
    requests:
      cpu: 10m
      memory: 1Gi
```

## 2) Best-Effort（尽力而为、不太可靠的）

如果 Pod 中所有容器都未定义资源配置（Requests 和 Limits 都未定义），那么该 Pod 的 QoS 级别就是 Best-Effort。

例如下面这个 Pod 定义：

```
containers:
  name: foo
    resources:
  name: bar
    resources:
```

## 3) Burstable（弹性波动、较可靠的）

当一个 Pod 既不是 Guaranteed 级别的，也不是 Best-Effort 级别的时，该 Pod 的 QoS 级别就是 Burstable。Burstable 级别的 Pod 包括两种情况。第 1 种情况是：Pod 中的一部分容器在一种或多种资源类型的资源配置中，定义了 Requests 值和 Limits 值（都不为 0），且 Requests 值小于 Limits 值；第 2 种情况是：Pod 中的一部分容器未定义资源配置（Requests 和 Limits 都未定义）。注意：容器未定义 Limits 时，Limits 值默认等于节点资源容量上限。

下面几个例子中的 Pod 的 QoS 等级都是 Burstable。

（1）容器 foo 的 CPU Requests 不等于 Limits：

```
containers:
  name: foo
    resources:
      limits:
        cpu: 10m
        memory: 1Gi
      requests:
        cpu: 5m
        memory: 1Gi
  name: bar
    resources:
```

```

limits:
  cpu: 10m
  memory: 1Gi
requests:
  cpu: 10m
  memory: 1Gi

```

(2) 容器 **bar** 未定义资源配置而容器 **foo** 定义了资源配置:

```

containers:
  name: foo
  resources:
    limits:
      cpu: 10m
      memory: 1Gi
    requests:
      cpu: 10m
      memory: 1Gi
  name: bar

```

(3) 容器 **foo** 未定义 CPU，而容器 **bar** 未定义内存:

```

containers:
  name: foo
  resources:
    limits:
      memory: 1Gi
  name: bar
  resources:
    limits:
      cpu: 100m

```

(4) 容器 **bar** 未定义资源配置，而容器 **foo** 未定义 Limits 值:

```

containers:
  name: foo
  resources:
    requests:
      cpu: 10m
      memory: 1Gi
  name: bar

```

#### 4) Kubernetes QoS 的工作特点

Pod 的 CPU Requests 无法得到满足(比如节点的系统级任务占用过多的 CPU 导致无法分配给足够的 CPU 给容器使用)时，容器得到的 CPU 会被压缩限流。

内存由于是不可压缩资源，所以针对内存资源紧缺的情况，将按照以下逻辑进行处理。

(1) Best-Effort Pod 的优先级最低，这类 Pod 中运行的进程会在系统内存紧缺时被第一优先

“杀死”。当然，从另外一个角度来看，Best-Effort Pod 由于没有设置资源 Limits，所以在资源充足时，它们可以充分地使用所有的闲置资源。

（2）Burstable Pod 的优先级居中，这类 Pod 初始时会分配较少的可靠资源，但可以按需申请更多的资源。当然，如果整个系统内存紧缺，而又没有 Best-Effort 容器可以被“杀死”以释放资源，则这类 Pod 中的进程可能会被“杀死”。

（3）Guaranteed Pod 的优先级最高，而且一般情况下这类 Pod 只要不超过其资源 Limits 的限制就不会被“杀死”。当然，如果整个系统内存紧缺，而又没有其他更低优先级的容器可以被“杀死”以释放资源，这类 Pod 中的进程也可能被“杀死”。

5) OOM 计分系统

OOM（Out Of Memory）计分规则包括如下内容。

- ◎ OOM 计分是一个进程消耗内存存在系统中占的百分比中不含百分号的数字的值乘以 10 的结果，这个结果是进程 OOM 基础分；将进程 OOM 基础分的分值再加上这个进程的 OOM 分数调整值 OOM\_SCORE\_ADJ 的值作为进程 OOM 最终分值（除 root 启动的进程外）。在系统发生 OOM 时，OOM Killer 会优先杀掉 OOM 计分更高的进程。
- ◎ 进程的 OOM 计分的基本分数值范围是 0～1000，如果 A 进程的调整值 OOM\_SCORE\_ADJ 减去 B 进程的调整值的结果大于 1000，那么 A 进程的 OOM 计分最终值必然大于 B 进程，A 进程会比 B 进程优先被杀死。
- ◎ 不论调整值 OOM\_SCORE\_ADJ 为多少，任何进程的最终分值范围也是 0～1000。

在 Kubernetes，不同 QoS 的 OOM 计分调整值规则如表 5.1 所示。

表 5.1 不同 QoS 的 OOM 计分调整值

QoS 等级	oom_score_adj
Guaranteed	-998
BestEffort	1000
Burstable	$\min(\max(2, 1000 - (1000 * \text{memoryRequestBytes}) / \text{machineMemoryCapacityBytes}), 999)$

- ◎ Best-effort Pod 设置 OOM\_SCORE\_ADJ 调整值为 1000，因此 Best-effort Pod 中的容器里面的所有进程的 OOM 最终分肯定是 1000。
- ◎ Guaranteed Pod 设置 OOM\_SCORE\_ADJ 调整值为-998，因此 Guaranteed Pod 中的容器里面的所有进程的 OOM 最终分一般为 0 或者 1（因为基础分不可能为 1000）。
- ◎ Burstable Pod 规则分情况说明：如果 Burstable Pod 的内存 Requests 超过了系统可用内存的 99.8%，那么这个 Pod 的 OOM\_SCORE\_ADJ 调整值固定为 2；否则，设置

OOM\_SCORE\_ADJ 调整值为  $1000 - 10$  (内存 Requests 占系统可用内存的百分比的无百分号的数字部分的值), 而如果内存 Requests 为 0, 那么 OOM\_SCORE\_ADJ 调整值固定为 999。这样的规则能确保 OOM\_SCORE\_ADJ 调整值的范围为 2~999, 而 Burstable Pod 中所有进程的 OOM 最终分数范围为 2~1000。Burstable Pod 进程的 OOM 最终分数始终大于 Guaranteed Pod 的进程得分, 因此它们会被优先“杀死”。如果一个 Burstable Pod 使用的内存比它的内存 Requests 少, 那么可以肯定的是它的所有进程的 OOM 最终分数会小于 1000, 此时能确保它的优先级高于 Best-effort Pod。如果一个 Burstable Pod 的某个容器中某个进程使用的内存比容器的 request 值高, 那么这个进程的 OOM 最终分数会是 1000, 否则它的 OOM 最终分会小于 1000。假设下面容器中有一个占用内存非常大的进程, 那么当一个使用内存超过其 Requests 的 Burstable Pod 与另外一个使用内存少于其 Requests 的 Burstable Pod 发生内存竞争冲突时, 前者的进程会被系统“杀掉”。如果一个 Burstable Pod 内部有多个进程的多个容器发生内存竞争冲突, 那么此时 OOM 评分只能作为参考, 不能保证完全按照资源配置的定义来执行 OOM Kill。

OOM 还有一些特殊的计分规则, 如下所述。

- ◎ kubelet 进程和 Docker 进程的调整值 OOM\_SCORE\_ADJ 为-998。
- ◎ 如果配置进程调整值 OOM\_SCORE\_ADJ 为-999, 那么这类进程不会被 OOM Killer “杀掉”。

## 6) QoS 的演进

目前 Kubernetes 基于 QoS 的超用机制日趋完善, 但还有一些问题需要解决。

## 7) 内存 Swap 的支持

当前的 QoS 策略都是假定主机不启用内存 Swap。如果主机启用了 Swap, 那么上面的 QoS 策略可能会失效。举例说明: 两个 Guaranteed Pod 都刚好达到了内存 Limits, 那么由于内存 Swap 机制, 它们还可以继续申请使用更多的内存。如果 Swap 空间不足, 那么最终这两个 Pod 中的进程可能会被“杀掉”。由于 Kubernetes 和 Docker 尚不支持内存 Swap 空间的隔离机制, 所以这一功能暂时还未实现。

## 8) 更丰富的 QoS 策略

当前的 QoS 策略都是基于 Pod 的资源配置 (Requests 和 Limits) 来定义的, 而资源配置本身又承担着对 Pod 资源管理和限制的功能。两种不同维度的功能使用同一个参数来配置, 可能会导致某些复杂需求无法满足, 比如当前 Kubernetes 无法支持弹性的、高优先级的 Pod。自定义 QoS 优先级能提供更大的灵活性, 完美地实现各类需求, 但同时会引入更高的复杂性, 而且过于灵活的设置会给予用户过高的权限, 对系统管理也提出了更大的挑战。

#### 4. 资源的配额管理（Resource Quotas）

如果一个 Kubernetes 集群被多个用户或者多个团队共享使用，那么就需要考虑共享时对资源公平使用的问题，因为某个用户可能会使用超过基于公平原则分配给其的资源量。

资源配额（Resource Quotas）就是解决这个问题的工具。通过 ResourceQuota 对象，我们可以定义一项资源配额，这个资源配额可以为每一个命名空间（namespace）提供一个总体的资源使用的限制：它可以限制命名空间中某种类型的对象的总数目上限，也可以设置命名空间中 Pod 可以使用到的计算资源的总上限。

典型的资源配额（Resource Quotas）使用方式如下。

- ◎ 不同的团队工作在不同的命名空间下，目前这个是非约束性的，未来版本中可能会通过 ACLs（访问控制列表 Access Control List）的方式来实现强制性约束。
- ◎ 集群管理员为集群中的每个命名空间创建一个或者多个资源配额项。
- ◎ 当用户在命名空间中使用资源（创建 Pod 或者 Service 等）时，Kubernetes 的配额系统会统计、监控和检查资源用量，以确保使用的资源用量没有超过资源配额的配置。
- ◎ 如果创建或者更新应用时，资源使用超过了某项资源配额的限制，那么创建或者更新的请求会报错（HTTP 403 Forbidden），错误信息给出详细的出错原因说明。
- ◎ 如果命名空间中的计算资源（CPU 和内存）的资源配额启用，那么用户必须为相应的资源类型设置 Requests 或 Limits；否则配额系统可能会直接拒绝 Pod 的创建。这里可以使用 LimitRange 机制来为没有配置资源的 Pod 提供默认资源配置。

下面的例子展示了一个非常适合使用资源配额来做资源控制管理的场景。

- ◎ 集群共有 32GB 内存和 16 CPU，两个小组，A 小组使用 20GB 内存和 10 CPU，B 小组使用 10GB 内存和 2 CPU，剩下的 2GB 内存和 2 CPU 作为预留。
- ◎ 在名为 testing 的命名空间中，限制使用 1 CPU 和 1GB 内存；在名为 production 的命名空间中，资源使用不受限制。

在使用资源配额时，需要注意以下两点。

- ◎ 如果集群中总的可用资源小于各命名空间中资源配额的总和，那么可能会导致资源竞争。资源竞争时，Kubernetes 系统使用先到先得的原则。
- ◎ 不管是资源竞争还是配额的修改都不会影响到已经创建的资源使用对象。

##### 1) 在 Master 中开启资源配额选型

资源配额可以通过在 kube-apiserver 的 --admission-control= 参数值中添加 ResourceQuota 参数进行开启。如果某个命名空间的定义中存在 ResourceQuota，那么对于该命名空间而言，资源配



额就是开启的。一个命名空间可以有多个 ResourceQuota 配置项。

### （1）计算资源配额（Compute Resource Quota）

资源配额可以限制一个命名空间中所有 Pod 的计算资源的总和。表 5.2 列出了目前 ResourceQuota 支持限制的計算资源类型。

表 5.2 ResourceQuota 支持限制的計算资源类型

资源名称	说明
Cpu	所有非终止状态的 Pod，CPU Requests 的总和不能超过该值
limits.cpu	所有非终止状态的 Pod，CPU Limits 的总和不能超过该值
limits.memory	所有非终止状态的 Pod，内存 Limits 的总和不能超过该值
Memory	所有非终止状态的 Pod，内存 Requests 的总和不能超过该值
requests.cpu	所有非终止状态的 Pod，CPU Requests 的总和不能超过该值
requests.memory	所有非终止状态的 Pod，内存 Requests 的总和不能超过该值

### （2）对象数量配额（Object Count Quota）

指定类型的对象数量可以被限制。表 5.3 列出了 ResourceQuota 支持限制的对象类型。

表 5.3 ResourceQuota 支持限制的对象类型

资源名称	说明
Configmaps	在该命名空间中，能存在的 ConfigMap 的总数上限
Persistentvolumeclaims	在该命名空间中，能存在的持久卷的总数上限
Pods	在该命名空间中，能存在的非终止状态 Pod 的总数上限。Pod 终止状态等价于 Pod 的 status.phase 状态值为 Failed 或者 Succeed is true
Replicationcontrollers	在该命名空间中，能存在的 RC 的总数上限
Resourcequotas	在该命名空间中，能存在的资源配额项（ResourcesQuota）的总数上限
Services	在该命名空间中，能存在的 service 的总数上限
services.loadbalancers	在该命名空间中，能存在的负载均衡（LoadBalancer）的总数上限
services.nodeports	在该命名空间中，能存在的 NodePort 的总数上限
Secrets	在该命名空间中，能存在的 Secret 的总数上限

例如我们可以通过资源配额来限制命名空间中能创建的 Pod 的最大数量。这种设置可以防止某些用户大量创建 Pod 而迅速耗尽整个集群的 Pod IP 和计算资源。

### 2）配额的作用域（Quota Scopes）

每项资源配额都可以单独配置一组作用域，配置了作用域的资源配额只会对符合其作用域的资源使用进行计量和限制，作用域范围内的且超过了资源配额请求都会报验证错。表 5.4 列出了 ResourceQuota 的 4 种作用域。

表 5.4 ResourceQuota 的 4 种作用域

作用域	说明
Terminating	匹配所有 spec.activeDeadlineSeconds >= 0 的 Pod
NotTerminating	匹配所有 spec.activeDeadlineSeconds 是 nil 的 Pod
BestEffort	匹配所有 QoS 为 Best-Effort 的 Pod
NotBestEffort	匹配所有 QoS 不是 Best-Effort 的 Pod

其中，BestEffort 作用域可以限定资源配额来追踪 pods 资源的使用，Terminating、NotTerminating 和 NotBestEffort 这三种作用域可以限定资源配额来追踪以下资源的使用。

- cpu
- limits.cpu
- limits.memory
- memory
- pods
- requests.cpu
- requests.memory

3) 在资源配额（ResourceQuota）中设置 Requests 和 Limits

资源配额也可以设置 Requests 和 Limits。

如果资源配额中指定了 requests.cpu 或 requests.memory，那么它会强制要求每一个容器都必须配置自己的 CPU Requests 或 CPU Limits（可使用 LimitRange 提供的默认值）。

同理，如果资源配额中指定了 limits.cpu 或 limits.memory，那么它也会强制要求每一个容器都必须配置自己的内存 Requests 或内存 Limits（可使用 LimitRange 提供的默认值）。

4) 资源配额（ResourceQuota）的定义

下面通过几个例子对资源配额进行设置和应用。

与 LimitRange 相似，ResourceQuota 也设置在 namespace 中。创建名为 myspace 的 namespace：

```
$ kubectl create namespace myspace
namespace "myspace" created
```

创建 ResourceQuota 配置文件 compute-resources.yaml，用于设置计算资源的配额：

```
apiVersion: v1
kind: ResourceQuota
metadata:
  name: compute-resources
spec:
```

```

hard:
  pods: "4"
  requests.cpu: "1"
  requests.memory: 1Gi
  limits.cpu: "2"
  limits.memory: 2Gi

```

创建该项资源配额:

```
$ kubectl create -f compute-resources.yaml --namespace=myspace
resourcequota "compute-resources" created
```

创建另一个名为 object-counts.yaml 的文件, 用于设置对象数量的配额:

```

apiVersion: v1
kind: ResourceQuota
metadata:
  name: object-counts
spec:
  hard:
    configmaps: "10"
    persistentvolumeclaims: "4"
    replicationcontrollers: "20"
    secrets: "10"
    services: "10"
    services.loadbalancers: "2"

```

创建该 ResourceQuota:

```
$ kubectl create -f object-counts.yaml --namespace=myspace
resourcequota "object-counts" created
```

查看各 ResourceQuota 的详细信息:

```
$ kubectl describe quota compute-resources --namespace=myspace
```

```

Name:                compute-resources
Namespace:           myspace
Resource              Used Hard
-----
limits.cpu            0    2
limits.memory         0    2Gi
pods                  0    4
requests.cpu          0    1
requests.memory       0    1Gi

```

```
$ kubectl describe quota object-counts --namespace=myspace
```

```

Name:                object-counts
Namespace:           myspace
Resource              Used   Hard
-----
configmaps            0     10

```

<code>persistentvolumeclaims</code>	0	4
<code>replicationcontrollers</code>	0	20
<code>secrets</code>	1	10
<code>services</code>	0	10
<code>services.loadbalancers</code>	0	2

### 5) 资源配额与集群资源总量的关系

资源配额与集群资源总量是完全独立的。资源配额是通过绝对的单位来配置的：这也就意味着如果集群中新添加了节点，那么资源配额不会自动更新，而该资源配额所对应的命名空间下对象也不能自动地增加资源上限。

在某些情况下，我们可能希望资源配额能支持更复杂的策略，如下所述。

- ◎ 对于不同的租户，按照比例划分整个集群的资源。
- ◎ 允许每个租户都能按照需要来提高资源用量，但是有一个较宽容的限制，以防止意外的资源耗尽情况发生。
- ◎ 探测某个命名空间的需求，添加物理节点并扩大资源配额值。

这些策略可以通过将资源配额作为一个控制模块、手动编写一个控制器（controller）来监控资源使用情况，并调整命名空间上的资源配额的方式来实现。

资源配额将整个集群中的资源总量做了一个静态的划分，但它并没有对集群中的节点（Node）做任何限制：不同命名空间中的 Pod 仍然可以运行到同一个节点上。

## 5. ResourceQuota 和 LimitRange 实践指南

根据前面对资源管理的介绍，这里将通过一个完整的例子来说明如何通过资源配额和资源范围范围的配合来控制一个命名空间的资源使用。

集群管理员根据集群用户数量来调整集群配置，以达到如下目的：能控制特定命名空间中的资源使用量，最终实现集群的公平使用和成本的控制。

需要实现的功能如下。

- ◎ 限制运行状态的 Pod 的计算资源用量。
- ◎ 限制持久存储卷的数量以控制对存储的访问。
- ◎ 限制负载均衡器的数量以控制成本。
- ◎ 防止滥用网络端口这类稀缺资源。
- ◎ 提供默认的计算资源 Requests 以便于系统做出更优化的调度。

## 1) 创建命名空间

创建名为 quota-example 的命名空间，namespace.yaml 文件的内容如下：

```
apiVersion: v1
kind: Namespace
metadata:
  name: quota-example

$ kubectl create -f namespace.yaml
namespace "quota-example" created
```

查看命名空间：

```
$ kubectl get namespaces
NAME             STATUS    AGE
default          Active   2m
kube-system      Active   2m
quota-example    Active   39s
```

## 2) 设置限定对象数目的资源配额

通过设置限定对象的数量资源配额，可以控制以下资源的数量：

- ◎ 持久存储卷；
- ◎ 负载均衡器；
- ◎ NodePort。

创建名为 object-counts 的 ResourceQuota：

**object-counts.yaml：**

```
apiVersion: v1
kind: ResourceQuota
metadata:
  name: object-counts
spec:
  hard:
    persistentvolumeclaims: "2"
    services.loadbalancers: "2"
    services.nodeports: "0"
```

```
$ kubectl create -f object-counts.yaml --namespace=quota-example
resourcequota "object-counts" created
```

配额系统会检测到资源项配额的创建，并且将会统计和限制该命名空间中的资源消耗。

查看该配额是否生效：

```
$ kubectl describe quota object-counts --namespace=quota-example
Name:                object-counts
```

```
Namespace:          quota-example
Resource            Used    Hard
-----
persistentvolumeclaims 0      2
services.loadbalancers 0      2
services.nodeports    0      0
```

至此，配额系统会自动阻止那些使资源用量超过资源配额限定值的请求。

### 3) 设置限定计算资源的资源配额

下面我们再来创建一项限定计算资源的资源配额，以限制该命名空间中的计算资源的使用总量。

创建名为 `compute-resources` 的 `ResourceQuota`：

```
apiVersion: v1
kind: ResourceQuota
metadata:
  name: compute-resources
spec:
  hard:
    pods: "4"
    requests.cpu: "1"
    requests.memory: 1Gi
    limits.cpu: "2"
    limits.memory: 2Gi
```

```
$ kubectl create -f compute-resources.yaml --namespace=quota-example
resourcequota "compute-resources" created
```

查看该配额是否生效：

```
$ kubectl describe quota compute-resources --namespace=quota-example
Name:          compute-resources
Namespace:     quota-example
Resource       Used Hard
-----
limits.cpu     0    2
limits.memory  0    2Gi
pods           0    4
requests.cpu   0    1
requests.memory 0    1Gi
```

配额系统会自动防止该命名空间下同时拥有超过 4 个非“终止态”的 Pod。此外，由于该项资源配额限制了 CPU 和内存的 Limits 和 Requests 的总量，因此会强制要求该命名空间下的所有容器都必须显示地定义 CPU 和内存的 Limits 和 Requests（可使用默认值，Requests 默认等于 Limits）。

#### 4) 配置默认 Requests 和 Limits

在命名空间已经配置了限定计算资源的资源配额的情况下，如果尝试在该命名空间下创建一个不指定 Requests 和 Limits 的 Pod，那么 Pod 的创建可能会失败。下面是一个失败的例子。

创建一个 Nginx 的 Deployment:

```
$ kubectl run nginx --image=nginx --replicas=1 --namespace=quota-example
deployment "nginx" created
```

查看创建的 Pod，会发现 Pod 没有创建成功:

```
$ kubectl get pods --namespace=quota-example
```

再查看一下 Deployment 的详细信息:

```
$ kubectl describe deployment nginx --namespace=quota-example
Name:          nginx
Namespace:     quota-example
CreationTimestamp:  Mon, 06 Jun 2016 16:11:37 -0400
Labels:        run=nginx
Selector:      run=nginx
Replicas:      0 updated | 1 total | 0 available | 1 unavailable
StrategyType:  RollingUpdate
MinReadySeconds: 0
RollingUpdateStrategy: 1 max unavailable, 1 max surge
OldReplicaSets:  <none>
NewReplicaSet:   nginx-3137573019 (0/1 replicas created)
.....
```

本 Deployment 尝试创建一个 Pod，但是失败了，查看其中 ReplicaSet 的详细信息:

```
$ kubectl describe rs nginx-3137573019 --namespace=quota-example
Name:          nginx-3137573019
Namespace:     quota-example
Image(s):      nginx
Selector:      pod-template-hash=3137573019,run=nginx
Labels:        pod-template-hash=3137573019
                run=nginx
Replicas:      0 current / 1 desired
Pods Status:   0 Running / 0 Waiting / 0 Succeeded / 0 Failed
No volumes.
Events:
  FirstSeen    LastSeen    Count  From              SubobjectPath  Type
Reason         Message
-----
4m             7s         11     {replicaset-controller }      Warning
FailedCreate Error creating: pods "nginx-3137573019-" is forbidden: Failed quota:
compute-resources: must specify limits.cpu,limits.memory,requests.cpu,requests.
memory
```

可以看到 Pod 创建失败的原因：**Master 拒绝了这个 ReplicaSet 创建 Pod，因为这个 Pod 中没有指定 CPU 和内存的 Requests 和 Limits。**

为了避免这种失败，我们可以使用 **LimitRange** 来为这个命名空间下的所有 Pod 提供一个资源配置的默认值。下面的例子展示了如何为这个命名空间添加一个指定默认资源配置的 **LimitRange**。

创建一个名为 **limits** 的 **LimitRange**：

```
limits.yaml:
apiVersion: v1
kind: LimitRange
metadata:
  name: limits
spec:
  limits:
  - default:
      cpu: 200m
      memory: 512Mi
    defaultRequest:
      cpu: 100m
      memory: 256Mi
    type: Container
```

```
$ kubectl create -f limits.yaml --namespace=quota-example
limitrange "limits" created
```

```
$ kubectl describe limits limits --namespace=quota-example
Name:          limits
Namespace:     quota-example
Type          Resource  Min  Max  Default Request  Default Limit  Max Limit/Request
Ratio
-----
Container memory -    -    256Mi          512Mi          -
Container cpu   -    -    100m           200m           -
```

**LimitRange** 创建成功后，用户在该命名空间下的创建未指定资源配置的 Pod 的请求时，系统会自动为该 Pod 设置默认的资源配置。

例如，每个新建的未指定资源配置的 Pod 都等价于使用下面的资源配置：

```
$ kubectl run nginx \
--image=nginx \
--replicas=1 \
--requests=cpu=100m,memory=256Mi \
--limits=cpu=200m,memory=512Mi \
--namespace=quota-example
```



至此，我们已经为该命名空间配置好了默认的计算资源，我们的 `ReplicaSet` 应该能够创建 Pod 了。查看一下，创建 Pod 成功了：

```
$ kubectl get pods --namespace=quota-example
NAME                                READY    STATUS    RESTARTS   AGE
nginx-3137573019-fvrig             1/1      Running   0           6m
```

接下来，还可以随时查看资源配额的使用情况：

```
$ kubectl describe quota --namespace=quota-example
Name:                compute-resources
Namespace:           quota-example
Resource             Used    Hard
-----
limits.cpu           200m    2
limits.memory        512Mi   2Gi
pods                 1        4
requests.cpu         100m    1
requests.memory      256Mi   1Gi
```

```
Name:                object-counts
Namespace:           quota-example
Resource             Used    Hard
-----
persistentvolumeclaims 0        2
services.loadbalancers 0        2
services.nodeports     0        0
```

可以看到每个 Pod 创建时都会消耗掉指定的资源量，而这些使用量都会被 Kubernetes 准确地跟踪、监控和管理。

### 5) 指定资源配额的作用域

假设我们并不想为某个命名空间配置默认的计算资源配额，而是希望限定在命名空间内运行的 QoS 为 `BestEffort` 的 Pod 总数，例如将集群中的部分资源用来运行 QoS 为非 `BestEffort` 的服务，而将闲置的资源用来运行 QoS 为 `BestEffort` 的服务，即可避免集群的所有资源仅被大量的 `BestEffort` Pod 耗尽。这可以通过创建两个资源配额（`ResourceQuota`）来实现。

首先创建一个名为 `quota-scopes` 的命名空间：

```
$ kubectl create namespace quota-scopes
namespace "quota-scopes" created
```

创建一个名为 `best-effort` 的 `ResourceQuota`，指定 Scope 为 `BestEffort`：

```
apiVersion: v1
kind: ResourceQuota
metadata:
```

```
name: best-effort
spec:
  hard:
    pods: "10"
  scopes:
  - BestEffort
```

```
$ kubectl create -f best-effort.yaml --namespace=quota-scopes
resourcequota "best-effort" created
```

再创建一个名为 not-best-effort 的 ResourceQuota，指定 Scope 为 NotBestEffort:

```
apiVersion: v1
kind: ResourceQuota
metadata:
  name: not-best-effort
spec:
  hard:
    pods: "4"
    requests.cpu: "1"
    requests.memory: 1Gi
    limits.cpu: "2"
    limits.memory: 2Gi
  scopes:
  - NotBestEffort
```

```
$ kubectl create -f not-best-effort.yaml --namespace=quota-scopes
resourcequota "not-best-effort" created
```

查看创建成功的 ResourceQuota:

```
$ kubectl describe quota --namespace=quota-scopes
```

```
Name:          best-effort
Namespace:     quota-scopes
Scopes:        BestEffort
* Matches all pods that have best effort quality of service.
Resource      Used  Hard
-----
pods          0    10
```

```
Name:          not-best-effort
Namespace:     quota-scopes
Scopes:        NotBestEffort
* Matches all pods that do not have best effort quality of service.
Resource      Used  Hard
-----
limits.cpu    0    2
limits.memory 0    2Gi
pods          0    4
```

```
requests.cpu      0      1
requests.memory   0      1Gi
```

之后，对于没有配置 Requests 的 Pod 将会被名为 best-effort 的 ResourceQuota 所限制；而配置了 Requests 的 Pod 会被名为 not-best-effort 的 ResourceQuota 所限制。

创建两个 Deployment：

```
$ kubectl run best-effort-nginx --image=nginx --replicas=8
--namespace=quota-scopes
deployment "best-effort-nginx" created
```

```
$ kubectl run not-best-effort-nginx \
  --image=nginx \
  --replicas=2 \
  --requests=cpu=100m,memory=256Mi \
  --limits=cpu=200m,memory=512Mi \
  --namespace=quota-scopes
deployment "not-best-effort-nginx" created
```

名为 best-effort-nginx 的 Deployment 因为没有配置 Requests 和 Limits，所以它的 QoS 级别为 BestEffort，因此它的创建过程由 best-effort 资源配额项来限制，而 not-best-effort 资源配额项不会对它进行限制。best-effort 资源配额项没有限制 Requests 和 Limits，因此 best-effort-nginx Deployment 可以成功地创建 8 个 Pod。

名为 not-best-effort-nginx 的 Deployment 因为配置了 Requests 和 Limits，且二者不相等，所以它的 QoS 级别为 Burstable，因此它的创建过程由 not-best-effort 资源配额项来限制，而 best-effort 资源配额项不会对它进行限制。not-best-effort 资源配额项限制了 Pod 的 Requests 和 Limits 的总上限，not-best-effort-nginx Deployment 并没有超过这个上限，所以可以成功地创建两个 Pod。

查看已经创建的 Pod：

```
$ kubectl get pods --namespace=quota-scopes
```

NAME	READY	STATUS	RESTARTS	AGE
best-effort-nginx-3488455095-2qb41	1/1	Running	0	51s
best-effort-nginx-3488455095-3go7n	1/1	Running	0	51s
best-effort-nginx-3488455095-9o2xg	1/1	Running	0	51s
best-effort-nginx-3488455095-eyg40	1/1	Running	0	51s
best-effort-nginx-3488455095-gcs3v	1/1	Running	0	51s
best-effort-nginx-3488455095-rq8p1	1/1	Running	0	51s
best-effort-nginx-3488455095-udhhd	1/1	Running	0	51s
best-effort-nginx-3488455095-zmk12	1/1	Running	0	51s
not-best-effort-nginx-2204666826-7sl61	1/1	Running	0	23s
not-best-effort-nginx-2204666826-ke746	1/1	Running	0	23s

可以看到 10 个 Pod 都创建成功。

再看一下两个资源配额项的使用情况：

```
$ kubectl describe quota --namespace=quota-scopes
Name:          best-effort
Namespace:     quota-scopes
Scopes:        BestEffort
* Matches all pods that have best effort quality of service.
Resource      Used  Hard
-----
pods          8    10

Name:          not-best-effort
Namespace:     quota-scopes
Scopes:        NotBestEffort
* Matches all pods that do not have best effort quality of service.
Resource      Used  Hard
-----
limits.cpu    400m  2
limits.memory 1Gi   2Gi
pods          2     4
requests.cpu   200m  1
requests.memory 512Mi 1Gi
```

可以看到 **best-effort** 资源配额项已经统计到了 **best-effort-nginx Deployment** 中创建的 8 个 Pod 的资源使用信息，而 **not-best-effort** 资源配额项也统计到了 **not-best-effort-nginx Deployment** 中创建的两个 Pod 的资源使用信息。

通过这个例子我们可以看到：资源配额的作用域（**Scopes**）提供了一种将资源集合分割的机制，这种机制使得集群管理员可以更加方便地监控和限制不同类型对象对于各类资源的使用，同时能为资源分配和限制提供更大的灵活度和便利性。

6. 资源管理总结

Kubernetes 中的资源管理的基础是容器和 Pod 的资源配置（**Requests** 和 **Limits**）。容器的资源配置（**Requests** 和 **Limits**）指定了容器请求的资源 and 容器能使用的资源上限，而 Pod 的资源配置则是 Pod 中所有容器的资源配置总和的上限。

通过资源配额（**Resource Quota**）机制，我们可以对命名空间下所有 Pod 使用资源的总量进行限制，也可以对这个命名空间中指定类型的对象的数量进行限制。使用作用域可以让资源配额只对符合特定范围的对象加以限制，因此作用域（**Scopes**）机制可以使资源配额的策略更加丰富灵活。

如果我们需要对用户的 Pod 或容器的资源配置做更多的限制，则我们可以使用资源配置范围（**LimitRange**）来达到这个目的。**LimitRange** 可以有效地限制 Pod 和容器的资源配置的最大、

最小范围，也可以限制 Pod 和容器的 Limits 与 Requests 的最大比例上限，此外 LimitRange 还可以为 Pod 中的容器提供默认的资源配置。

Kubernetes 基于 Pod 的资源配置（Requests 和 Limits）实现了资源服务质量（QoS）。不同 QoS 级别的 Pod 在系统中拥有不同的优先级：高优先级的 Pod 具有更高的可靠性，可以用于运行可靠性要求较高的服务；而低优先级的 Pod 可以实现集群资源的超售，能有效地提高集群资源利用率。

上面的多种机制共同组成了当前版本 Kubernetes 的资源管理体系。这个资源管理体系已经可以满足大部分资源管理的需求了。同时，Kubernetes 资源管理体系仍然在不停地发展和进化中，对于一些目前无法满足的更复杂、更个性化的需求，我们可以继续关注 Kubernetes 未来的发展和变化。

### 5.1.5 资源紧缺时的 Pod 驱逐机制

如何在系统硬件资源紧缺的情况下保证 Node 的稳定性，是 kubelet 需要解决的一个重要问题。尤其对于内存和磁盘这种不可压缩的资源，紧缺就意味着不稳定。下面对驱逐的策略、信号、阈值、监控频率和驱逐操作进行详细说明。

#### 1. 驱逐策略

kubelet 持续监控主机的资源使用情况，并尽量防止计算资源被耗尽。一旦出现资源紧缺的迹象，kubelet 就会主动终止一或多个 Pod 的运行，以回收紧缺的资源。当一个 Pod 被终止时，其中的容器会全部停止，Pod 的状态会被设置为 Failed。

#### 2. 驱逐信号

表 5.5 中提到了一些信号，kubelet 能够利用这些信号作为决策依据来触发驱逐行为。描述列中的内容来自于 kubelet Summary API。每个信号都支持整数或者百分比的表示方法。百分比的分母部分就是各个信号相关资源的总量。

表 5.5 驱逐信号及其描述

驱 逐 信 号	描 述
memory.available	memory.available := node.status.capacity[memory] - node.stats.memory.workingSet
nodefs.available	nodefs.available := node.stats.fs.available
nodefs.inodesFree	nodefs.inodesFree := node.stats.fs.inodesFree
imagefs.available	imagefs.available := node.stats.runtime.imagefs.available
imagefs.inodesFree	imagefs.inodesFree := node.stats.runtime.imagefs.inodesFree

`memory.available` 的值取自于 `cgroupfs`，而不是 `free -m` 命令，这是因为 `free -m` 不支持在容器内工作。如果用户使用了 `node allocatable` 功能，除了节点自身的内存需要判断，还需要利用 `cgroup` 根据用户 Pod 部分的情况进行判断。下面的脚本展示了 `kubelet` 计算 `memory.available` 的过程：

```
#!/bin/bash
#!/usr/bin/env bash

# This script reproduces what the kubelet does
# to calculate memory.available relative to root cgroup.

# current memory usage
memory_capacity_in_kb=$(cat /proc/meminfo | grep MemTotal | awk '{print $2}')
memory_capacity_in_bytes=$((memory_capacity_in_kb * 1024))
memory_usage_in_bytes=$(cat /sys/fs/cgroup/memory/memory.usage_in_bytes)
memory_total_inactive_file=$(cat /sys/fs/cgroup/memory/memory.stat | grep
total_inactive_file | awk '{print $2}')

memory_working_set=$memory_usage_in_bytes
if [ "$memory_working_set" -lt "$memory_total_inactive_file" ];
then
    memory_working_set=0
else
    memory_working_set=$((memory_usage_in_bytes - memory_total_inactive_file))
fi

memory_available_in_bytes=$((memory_capacity_in_bytes - memory_working_set))
memory_available_in_kb=$((memory_available_in_bytes / 1024))
memory_available_in_mb=$((memory_available_in_kb / 1024))

echo "memory.capacity_in_bytes $memory_capacity_in_bytes"
echo "memory.usage_in_bytes $memory_usage_in_bytes"
echo "memory.total_inactive_file $memory_total_inactive_file"
echo "memory.working_set $memory_working_set"
echo "memory.available_in_bytes $memory_available_in_bytes"
echo "memory.available_in_kb $memory_available_in_kb"
echo "memory.available_in_mb $memory_available_in_mb"
```

`kubelet` 假设 `inactive_file`（不活跃 LRU 列表中的 file-backed 内存，以字节为单位）在紧缺情况下可以回收，因此对其进行了排除。

`kubelet` 支持以下两种文件系统。

- (1) `nodefs`：保存 `kubelet` 的卷和守护进程日志等。
- (2) `imagefs`：在容器运行时用于保存镜像及可写入层。

kubelet 使用 cAdvisor 自动监控这些文件系统。kubelet 不关注其他文件系统，所有其他类型的配置，例如保存在独立文件系统上的卷和日志，都不被支持。

磁盘压力相关的资源回收机制正在逐渐被驱逐策略接管，未来将会停止对现有垃圾收集方式的支持。

### 3. 驱逐阈值

kubelet 可以定义驱逐阈值，一旦超出阈值，就会触发 kubelet 进行资源回收的操作。

阈值的定义方式为：

`<eviction-signal> <operator> <quantity>`

- ◎ 表 5.5 中列出了驱逐信号的名称。
- ◎ 当前仅支持一个 operator（运算符）：<（小于）。
- ◎ quantity 需要符合 Kubernetes 的数量表达方式，也可以用以%结尾的百分比表示。

例如，如果一个节点有 10Gi 内存，我们希望在可用内存不足 1Gi 时进行驱逐，就可以用下面任一方式来定义驱逐阈值。

- ◎ `memory.available<10%`。
- ◎ `memory.available<1Gi`。

驱逐阈值又可以通过软阈值和硬阈值两种方式进行设置。

#### 1) 驱逐软阈值

驱逐软阈值由一个驱逐阈值和一个管理员设定的宽限期共同定义。当系统资源消耗达到软阈值时，这一状况的持续时间在达到宽限期之前，kubelet 不会触发驱逐动作。如果没有定义宽限期，则 kubelet 会拒绝启动。

另外，可以定义终止 Pod 的宽限期。如果定义了这一宽限期，那么 kubelet 会使用 `pod.Spec.TerminationGracePeriodSeconds` 和最大宽限期这两个值之间较小的数值进行宽限，如果没有指定，则 kubelet 会立即杀掉 Pod。

软阈值的定义包括以下几个参数。

- ◎ `--eviction-soft`：描述驱逐阈值（例如 `memory.available<1.5Gi`），如果满足这一条件的持续时间超过宽限期，就会触发对 Pod 的驱逐动作。
- ◎ `--eviction-soft-grace-period`：驱逐宽限期（例如 `memory.available=1m30s`），用于定义达到软阈值之后持续时间超过多久才进行驱逐。

- ◎ `--eviction-max-pod-grace-period`: 在达到软阈值后，终止 Pod 的最大宽限期时间（单位为 s）。

2) 驱逐硬阈值

硬阈值没有宽限期，如果达到了硬阈值，则 kubelet 会立即杀掉 Pod 并进行资源回收。  
硬阈值的定义包括参数 `--eviction-hard`: 驱逐硬阈值，一旦达到阈值，就会触发对 Pod 的驱逐操作。

kubelet 的默认硬阈值定义如下：

```
--eviction-hard=memory.available<100Mi
```

4. 驱逐监控频率

kubelet 的 `--housekeeping-interval` 参数定义了一个时间间隔，kubelet 每隔一个这样的时间间隔就会对驱逐阈值进行评估。

5. 节点的状况

kubelet 会将一个或多个驱逐信号与节点的状况对应起来。

无论触发了硬阈值还是软阈值，kubelet 都会认为当前节点的压力太大，如表 5.6 所示为节点状况与驱逐信号的对应关系。

表 5.6 节点状况与驱逐信号的对应关系

节 点 状 况	驱 逐 信 号	描 述
MemoryPressure	memory.available	节点的可用内存达到了驱逐阈值
DiskPressure	nodefs.available, nodefs.inodesFree, imagefs.available, imagefs.inodesFree	节点的 root 文件系统或者镜像文件系统的可用空间达到了驱逐阈值

kubelet 会持续向 Master 报告节点状态的更新过程，这一频率由参数 `--node-status-update-frequency` 指定，默认为 10s。

6. 节点状况的抖动

如果一个节点的状况在软阈值的上下抖动，但是又没有超过宽限期，则将会导致该节点的相应状态在 True 和 False 之间不断变换，可能会对调度的决策过程产生负面影响。

要防止这种状况，则可以使用参数 `--eviction-pressure-transition-period`（脱离压力状态前需要等待的时间，默认值为 5m0s），为 kubelet 设置在脱离压力状态之前需要等待的时间。



这样一来，kubelet 在把压力状态设置为 False 之前，会确认在检测周期之内该节点没有达到驱逐阈值。

## 7. 回收 Node 级别的资源

如果达到了驱逐阈值，并且也过了宽限期，则 kubelet 会开始回收超出限量的资源，直到驱逐信号量回到阈值以内。

kubelet 在驱逐用户 Pod 之前，会尝试回收 Node 级别的资源。在观测到磁盘压力的情况下，基于服务器是否为容器运行时定义了独立的 imagefs，会导致不同的资源回收过程。

### 1) 有 Imagefs 的情况

(1) 如果 nodefs 文件系统达到了驱逐阈值，则 kubelet 会删除死掉的 Pod、容器来清理空间。

(2) 如果 imagefs 文件系统达到了驱逐阈值，则 kubelet 会删掉所有无用的镜像来清理空间。

### 2) 没有 Imagefs 的情况

如果 nodefs 文件系统到达了驱逐阈值，则 kubelet 会按照下面的顺序来清理空间。

(1) 删除死掉的 Pod、容器。

(2) 删除所有无用的镜像。

## 8. 驱逐用户的 Pod

如果 kubelet 无法通过节点级别的资源回收获取足够的资源，就会开始驱逐用户的 Pod。

kubelet 会按照下面的标准对 Pod 的驱逐行为进行判断。

- ◎ Pod 要求的服务质量。
- ◎ 根据 Pod 调度请求的被耗尽资源的消耗量。

接下来，kubelet 按照下面的顺序进行 Pod 的驱逐。

- ◎ BestEffort: 对紧缺资源消耗最多的 Pod 最先被驱逐。
- ◎ Burstable: 根据相对请求 (request) 来判断，对紧缺资源消耗最多的 Pod 最先被驱逐，如果没有 Pod 超出它们的请求，则策略会瞄准紧缺资源消耗量最大的 Pod。
- ◎ Guaranteed: 根据相对请求 (request) 来判断，对紧缺资源消耗最多的 Pod 最先被驱逐，如果没有 Pod 超出它们的请求，策略会瞄准紧缺资源消耗量最大的 Pod。

Guaranteed Pod 永远不会因为其他 Pod 的资源消费被驱逐。如果系统进程（例如 kubelet、docker、journald 等）消耗了超出 system-reserved 或者 kube-reserved 的资源，而这一节点上只运

行了 **Guaranteed Pod**，那么为了保证节点的稳定性并降低异常消费对其他 **Guaranteed Pod** 的影响，必须选择一个 **Guaranteed Pod** 进行驱逐。

本地磁盘是一种 **BestEffort** 资源。如有必要，**kubelet** 会在 **DiskPressure** 的情况下，对 **Pod** 进行驱逐以回收磁盘资源。**kubelet** 会按照 **QoS** 进行评估。如果 **kubelet** 判定缺乏 **inode** 资源，就会通过驱逐最低 **QoS** 的 **Pod** 的方式来回收 **inodes**。如果 **kubelet** 判定缺乏磁盘空间，就会在相同 **QoS** 的 **Pod** 中，选择消耗最多磁盘空间的 **Pod** 进行驱逐。下面针对有 **Imagefs** 和没有 **Imagefs** 的两种情况，说明 **kubelet** 在驱逐 **Pod** 时选择 **Pod** 的排序算法，然后按顺序对 **Pod** 进行驱逐。

#### 1) 有 **Imagefs** 的情况

如果 **nodefs** 触发了驱逐，则 **kubelet** 会根据 **nodefs** 的使用情况（以 **Pod** 中所有容器的本地卷和日志所占的空间进行计算）对 **Pod** 进行排序。

如果 **imagefs** 触发了驱逐，则 **kubelet** 会根据 **Pod** 中所有容器消耗的可写入层的使用空间进行排序。

#### 2) 没有 **Imagefs** 的情况

如果 **nodefs** 触发了驱逐，则 **kubelet** 会对各个 **Pod** 中所有容器的总体磁盘消耗（以本地卷+日志+所有容器的写入层所占的空间进行计算）进行排序。

### 9. 资源最少回收量

在某些场景下，驱逐 **Pod** 可能只回收了很少的资源，这就导致了 **kubelet** 反复触发驱逐阈值。另外，回收磁盘这样的资源，是需要消耗时间的。

要缓和这种状况，**kubelet** 可以对每种资源定义 **minimum-reclaim**。**kubelet** 一旦监测到了资源压力，就会试着回收不少于 **minimum-reclaim** 的资源数量，使得资源消耗量回到期望范围。

例如，可以配置 **--eviction-minimum-reclaim** 如下：

```
--eviction-hard=memory.available<500Mi,nodefs.available<1Gi,imagefs.available<100Gi  
--eviction-minimum-reclaim="memory.available=0Mi,nodefs.available=500Mi,imagefs.available=2Gi"
```

这样设置的效果如下。

- ◎ 当 **memory.available** 超过阈值触发了驱逐操作时，**kubelet** 会启动资源回收，并保证 **memory.available** 至少有 **500Mi**。
- ◎ 如果是 **nodefs.available** 超过阈值并触发了驱逐操作，则 **kubelet** 会恢复 **nodefs.available** 到至少 **1.5Gi**。

- ◎ 对于 `imagefs.available` 超过阈值并触发了驱逐操作的情况，`kubelet` 会保证 `imagefs.available` 恢复到最少 102Gi。

默认情况下，所有资源的 `eviction-minimum-reclaim` 为 0。

## 10. 节点资源紧缺情况下的系统行为

### 1) 调度器 (Scheduler) 的行为

在节点资源紧缺的情况下，节点会向 Master 报告这一状况。在 Master 上运行的调度器 (Scheduler) 以此为信号，不再继续向该节点调度新的 Pod。如表 5.7 所示为节点状况与调度行为的对应关系。

表 5.7 节点状况与调度行为的对应关系

节 点 状 况	调 度 行 为
MemoryPressure	不再调度新的 BestEffort Pod 到这个节点
DiskPressure	不再向这一节点调度 Pod

### 2) Node 的 OOM 行为

如果节点在 `kubelet` 能够回收内存之前，遭遇了系统的 OOM（内存不足），节点则依赖 `oom_killer` 的设置进行响应（OOM 评分系统详见 5.1.4 节的描述）。

`kubelet` 根据 Pod 的 QoS 为每个容器设置了一个 `oom_score_adj` 值，如表 5.8 所示。

表 5.8 `kubelet` 根据 Pod 的 QoS 为每个容器设置了一个 `oom_score_adj` 值

QoS 等级	oom_score_adj
Guaranteed	-998
BestEffort	1000
Burstable	$\min(\max(2, 1000 - (1000 * \text{memoryRequestBytes}) / \text{machineMemoryCapacityBytes}), 999)$

如果 `kubelet` 无法在系统 OOM 之前回收足够的内存，则 `oom_killer` 会根据内存使用比率来计算 `oom_score`，将得出的结果和 `oom_score_adj` 相加，最后得分最高的 Pod 会被首先驱逐。

这个策略的思路是，QoS 最低且相对于调度的 Request 来说消耗最多内存的 Pod 会被首先清除，来保障内存的回收。

与 Pod 驱逐不同，如果一个 Pod 的容器被 OOM 杀掉，是可能被 `kubelet` 根据 `RestartPolicy` 重启的。

### 3) 对 DaemonSet 类型 Pod 驱逐的考虑

通过 `DemonSet` 创建的 Pod 具有在节点上自动重启的特性，因此我们不希望 `kubelet` 驱

逐这种 Pod；然而 kubelet 目前并没有能力分辨 DaemonSet 的 Pod，所以无法单独为其制定驱逐策略，所以强烈建议不要在 DaemonSet 中创建 BestEffort 类型的 Pod，避免产生驱逐方面的问题。

10. 可调度的资源和驱逐策略最佳实践

假设一个集群的资源管理需求如下。

- ◎ 节点内存容量：10Gi。
- ◎ 保留 10% 的内存给系统守护进程（内核、kubelet 等）。
- ◎ 在内存使用率达到 95% 时驱逐 Pod，以此降低系统压力并防止系统 OOM。

为了满足这些需求，kubelet 应该设置如下参数：

```
--eviction-hard=memory.available<500Mi
--system-reserved=memory=1.5Gi
```

这个配置方式中隐式地包括了这样一个设置：系统预留内存也包括了资源驱逐阈值。

如果内存占用超出这一设置，则说明要么有 Pod 占用了超过其 Request 的内存，要么就是系统使用了超过 500Mi 内存。

在这种设置下，节点一旦开始接近内存压力，调度器就不会向该节点部署 Pod，并且假定这些 Pod 使用的资源数量少于其请求（Request）的资源数量。

11. 将被弃用的垃圾回收的 kubelet 参数

因为磁盘相关的驱逐参数已经趋于成熟，所以表 5.9 的 kubelet 参数会被标记为弃用或被新的参数替换。

表 5.9 现有参数及新参数的对应关系

现 有 参 数	新 参 数
--image-gc-high-threshold	--eviction-hard 或 eviction-soft
--image-gc-low-threshold	--eviction-minimum-reclaim
--maximum-dead-containers	弃用
--maximum-dead-containers-per-container	弃用
--minimum-container-ttl-duration	弃用
--low-diskspace-threshold-mb	--eviction-hard 或 eviction-soft
--outofdisk-transition-frequency	--eviction-pressure-transition-period

## 12. 现阶段的问题

### 1) kubelet 无法及时观测到内存压力

kubelet 目前从 cAdvisor 定时获取内存使用状况的统计情况。如果内存使用在这个时间段内发生了快速增长，且 kubelet 无法观察到 MemoryPressure，则可能会触发 OOMKiller。Kubernetes 正在尝试将这一过程集成到 memcg 通知 API 中来减少这一延迟，而不是让内核首先发现这一情况。

如果用户不希望将内存用尽，而是作为考察是否过量使用内存的敏感度量指标，则一个较为可靠的处理方式就是设置驱逐阈值为大约 75%，这样就降低了发生 OOM 的几率，提高了驱逐的标准，有助于集群状态的平衡。

### 2) kubelet 可能会错误地驱逐更多的 Pod

这也是状态搜集存在时间差导致的。未来可能会通过按需获取根容器的统计信息来减少计算偏差 (<https://github.com/google/cadvisor/issues/1247>)。

## 5.1.6 Pod Disruption Budget (主动驱逐保护)

在 Kubernetes 集群运行的过程中，许多管理操作都可能对 Pod 进行主动地驱逐，“主动”一词意味着这一操作可以安全地延迟一段时间，目前主要针对以下两种场景。

- ◎ 节点的维护或升级时 (kubectl drain)。
- ◎ 对应用的自动缩容操作 (autoscaling down)。

未来，rescheduler 也可能执行这个操作。

作为对比，由于节点不可用 (Not Ready) 导致的 Pod 驱逐就不能称之为主动了。

对于主动驱逐的场景来说，应用如果能够保持存活 Pod 的数量，则将会非常有用。通过使用 PodDisruptionBudget，应用可以保证那些会主动移除 Pod 的集群操作永远不会同一时间停掉太多 Pod，从而导致服务中断或者服务降级等后果。例如一个 Node 要进行维护时，系统应该保证应用以不低于一定数量的 Pod 保障服务的正常运行。kubectl drain 操作将遵循 PodDisruptionBudget 的设定，如果在该节点上运行了属于同一服务的多个 Pod，则为了保证最少存活数量，系统将确保每终止一个 Pod 后，一定会在另一台健康的 Node 上启动新的 Pod 后，再继续终止下一个 Pod。

PodDisruptionBudget 资源对象在 Kubernetes v1.5 版本时升级为 Beta 版本，用于指定一个 Pod 集合在一段时间内存活的最小实例数量或者百分比。一个 PodDisruptionBudget 作用于一组被同一个控制器管理的 Pod，例如 ReplicaSet 或 RC。与通常的 Pod 删除不同，驱逐 Pod 的控制

器将使用/eviction 接口对 Pod 进行驱逐，如果这一主动驱逐行为违反了 PodDisruptionBudget 的约定，就会被 API Server 拒绝。

PodDisruptionBudget 本身无法真正保障指定数量/百分比的 Pod 的存活。例如一个节点中包含了目标 Pod 中的一个，如果节点故障，就会导致 Pod 数少于 minAvailable。PodDisruptionBudget 对象的保护作用仅仅针对主动驱逐的场景，而非所有场景。

对 PodDisruptionBudget 的定义包括如下两部分。

- ◎ **Label Selector:** 用于筛选被管理的 Pod。
- ◎ **minAvailable:** 指定驱逐过程需要保障的最少 Pod 数量。minAvailable 可以是一个数字，也可以是一个百分比，例如 100%就表示不允许进行主动驱逐。

PodDisruptionBudget 示例如下。

(1) 首先创建一个 Deployment，Pod 数量为 3 个：

```
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: nginx
  labels:
    name: nginx
spec:
  replicas: 3
  selector:
    matchLabels:
      name: nginx
  template:
    metadata:
      labels:
        name: nginx
    spec:
      containers:
        - name: nginx
          image: nginx
          ports:
            - containerPort: 80
              protocol: TCP
```

创建后通过 `kubectl get pods` 命令查看 Pod 的创建情况：

NAME	READY	STATUS	RESTARTS	AGE
nginx-1968750913-0k01k	1/1	Running	0	13m
nginx-1968750913-1dpcn	1/1	Running	0	19m
nginx-1968750913-n326r	1/1	Running	0	13m

(2) 接下来创建一个 PodDisruptionBudget 对象：

```

apiVersion: policy/v1beta1
kind: PodDisruptionBudget
metadata:
  name: nginx
spec:
  minAvailable: 3
  selector:
    matchLabels:
      name: nginx

```

PodDisruptionBudget 使用的是和 Deployment 一样的 Label Selector, 并且设置存活 Pod 数量不得少于 3 个。

(3) 主动驱逐验证。对 Pod 的主动驱逐操作将通过驱逐 API (/eviction) 来完成。这个 API 可以视为受策略控制的对 Pod 的 DELETE 操作。要实现一次主动驱逐(更准确地说是一个 eviction), 则需要 POST 一个 JSON 请求, 以 eviction.json 文件格式表示, 内容如下:

```

{
  "apiVersion": "policy/v1beta1",
  "kind": "Eviction",
  "metadata": {
    "name": "nginx-1968750913-0k01k",
    "namespace": "default"
  }
}

```

用 curl 命令执行 eviction 操作:

```

$ curl -v -H 'Content-type: application/json'
http://<k8s_master>/api/v1/namespaces/default/pods/nginx-1968750913-0k01k/eviction -d @eviction.json

```

由于 PodDisruptionBudget 设置存活 Pod 数量不能少于 3 个, 因此驱逐操作会失败, 返回的错误信息中会包含如下内容:

```
"message": "Cannot evict pod as it would violate the pod's disruption budget."
```

使用 kubectl get pods 查看 Pod 列表, 会看到 Pod 数量和名字都没有发生变化。

最后用 kubectl delete pdb nginx 命令删除 pdb 对象。

再次执行上文中的 curl 指令, 会执行成功。用 kubectl get pods 命令查看 Pod 列表, 会发现 Pod 数量虽然没有发生变化, 但是指定的 Pod 已经消失, 取而代之的是一个新的 Pod。

### 5.1.7 Kubernetes 集群的高可用部署方案

Kubernetes 作为容器应用的管理平台, 通过对 Pod 的运行状况进行监控, 并且根据主机或

容器失效的状态将新的 Pod 调度到其他 Node 上，实现了应用层的高可用性。针对 Kubernetes 集群，高可用性还应包含以下两个层面的考虑：etcd 数据存储的高可用性和 Kubernetes Master 组件的高可用性。

1. etcd 高可用部署

etcd 在整个 Kubernetes 集群中处于中心数据库的地位，为保证 Kubernetes 集群的高可用性，首先需要保证数据库不是单故障点。一方面，etcd 需要以集群的方式进行部署，以实现 etcd 数据存储的冗余、备份与高可用性；另一方面，etcd 存储的数据本身也应考虑使用可靠的存储设备。

etcd 集群的部署可以使用静态配置，也可以通过 etcd 提供的 REST API 在运行时动态添加、修改或删除集群中的成员。本节将对 etcd 集群的静态配置进行说明。关于动态修改的操作方法请参考 etcd 官方文档的说明。

首先，规划一个至少 3 台服务器（节点）的 etcd 集群，在每台服务器上安装好 etcd。

部署一个由 3 台服务器组成的 etcd 集群，其配置如表 5.10 所示，其集群部署实例如图 5.5 所示。

表 5.10 etcd 集群的配置

etcd 实例名称	IP 地址
etcd1	10.0.0.1
etcd2	10.0.0.2
etcd3	10.0.0.3

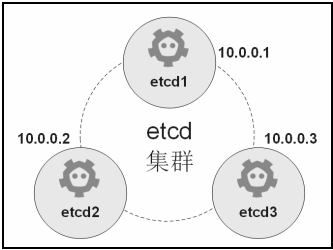


图 5.5 etcd 集群部署实例

然后修改每台服务器上 etcd 的配置文件/etc/etcd/etcd.conf。

以 etcd1 为创建集群的实例，需要将其 ETCD\_INITIAL\_CLUSTER\_STATE 设置为 “new”。etcd1 的完整配置如下：

```
# [member]
ETCD_NAME=etcd1           #etcd 实例名称
ETCD_DATA_DIR="/var/lib/etcd"   #etcd 数据保存目录
ETCD_LISTEN_CLIENT_URLS="http://10.0.0.1:2379,http://127.0.0.1:2379"   #供外部
客户端使用的 URL
ETCD_ADVERTISE_CLIENT_URLS="http://10.0.0.1:2379,http://127.0.0.1:2379"   #广
播给外部客户端使用的 URL
# [cluster]
ETCD_LISTEN_PEER_URLS="http://10.0.0.1:2380"   #集群内部通信使用的 URL
ETCD_INITIAL_ADVERTISE_PEER_URLS="http://10.0.0.1:2380"   #广播给集群内其他成员
访问的 URL
```



```
ETCD_INITIAL_CLUSTER="etcd1=http://10.0.0.1:2380,etcd2=http://10.0.0.2:2380,
etcd3=http://10.0.0.3:2380"      #初始集群成员列表
ETCD_INITIAL_CLUSTER_STATE="new"    #初始集群状态, new 为新建集群
ETCD_INITIAL_CLUSTER_TOKEN="etcd-cluster"  #集群名称
```

启动 etcd1 服务器上的 etcd 服务:

```
$ systemctl restart etcd
```

启动完成后, 就创建了一个名为 etcd-cluster 的集群。

etcd2 和 etcd3 为加入 etcd-cluster 集群的实例, 需要将其 ETCD\_INITIAL\_CLUSTER\_STATE 设置为 “exist”。etcd2 的完整配置如下 (etcd3 的配置略):

```
# [member]
ETCD_NAME=etcd2          #etcd 实例名称
ETCD_DATA_DIR="/var/lib/etcd"    #etcd 数据保存目录
ETCD_LISTEN_CLIENT_URLS="http://10.0.0.2:2379,http://127.0.0.1:2379"    #供外部
客户端使用的 URL
ETCD_ADVERTISE_CLIENT_URLS="http://10.0.0.2:2379,http://127.0.0.1:2379"    #广
播给外部客户端使用的 URL
#[cluster]
ETCD_LISTEN_PEER_URLS="http://10.0.0.2:2380"    #集群内部通信使用的 URL
ETCD_INITIAL_ADVERTISE_PEER_URLS="http://10.0.0.2:2380"    #广播给集群内其他成员
使用的 URL
ETCD_INITIAL_CLUSTER="etcd1=http://10.0.0.1:2380,etcd2=http://10.0.0.2:2380,
etcd3=http://10.0.0.3:2380"    #初始集群成员列表
ETCD_INITIAL_CLUSTER_STATE="new"    #初始集群状态, new 为新建集群
ETCD_INITIAL_CLUSTER_TOKEN="etcd-cluster"    #集群名称
```

启动 etcd2 和 etcd3 服务器上的 etcd 服务:

```
$ systemctl restart etcd
```

启动完成后, 在任意 etcd 节点执行 etcdctl cluster-health 命令来查询集群的运行状态:

```
$ etcdctl cluster-health
cluster is healthy
member ce2a822cea30bfca is healthy
member acda82balcf790fc is healthy
member eba209cd0012cd2 is healthy
```

在任意 etcd 节点上执行 etcdctl member list 命令来查询集群的成员列表:

```
$ etcdctl member list
ce2a822cea30bfca: name=default peerURLs=http://10.0.0.1:2380,http://127.0.0.1:
7001 clientURLs=http://10.0.0.1:2379,http://127.0.0.1:2379
acda82balcf790fc: name=default peerURLs=http://10.0.0.2:2380,http://127.0.0.1:
7001 clientURLs=http://10.0.0.2:2379,http://127.0.0.1:2379
eba209cd40012cd2: name=default peerURLs=http://10.0.0.3:2380,http://127.0.0.1:
7001 clientURLs=http://10.0.0.3:2379,http://127.0.0.1:2379
```

至此，一个 etcd 集群就创建成功了。

以 kube-apiserver 为例，将访问 etcd 集群的参数设置为：

```
--etcd-servers=http://10.0.0.1:2379,http://10.0.0.2:2379,http://10.0.0.3:2379
```

在 etcd 集群成功启动之后，如果需要对集群成员进行修改，则请参考官方文档的详细说明：

<https://github.com/coreos/etcd/blob/master/Documentation/runtime-configuration.md#cluster-reconfiguration-operations>。

对于 etcd 中需要保存的数据的可靠性，可以考虑使用 RAID 磁盘阵列、高性能存储设备、共享存储文件系统，或者使用云服务商提供的存储系统等来实现。

2. Master 高可用部署

在 Kubernetes 系统中，Master 服务扮演着总控中心的角色，主要的三个服务 kube-apiserver、kube-controller-manster 和 kube-scheduler 通过不断与工作节点上的 kubelet 和 kube-proxy 进行通信来维护整个集群的健康工作状态。如果 Master 的服务无法访问到某个 Node，则会将该 Node 标记为不可用，不再向其调度新建的 Pod。但对 Master 自身则需要进行额外的监控，使 Master 不成为集群的单故障点，所以对 Master 服务也需要进行高可用方式的部署。

以 Master 的 kube-apiserver、kube-controller-manster 和 kube-scheduler 三个服务作为一个部署单元，类似于 etcd 集群的典型部署配置。使用至少三台服务器安装 Master 服务，并且需要保证任何时候总有一套 Master 能够正常工作。图 5.6 展示了一种典型的部署方式。

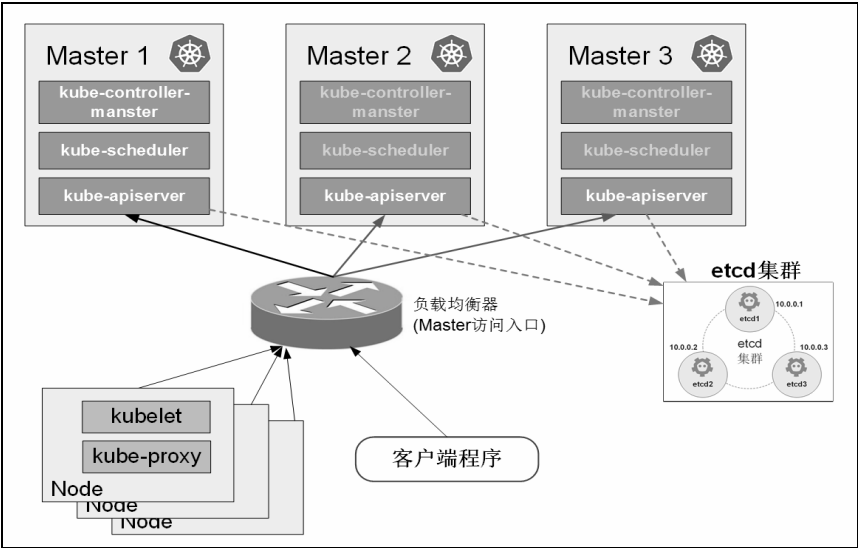


图 5.6 一种典型的部署方式

Kubernetes 建议 Master 的 3 个组件都以容器的形式启动，启动它们的基础工具是 kubelet，所以它们都将以 Static Pod 的形式启动并由 kubelet 进行监控和自动重启。而 kubelet 本身的高可用则通过操作系统来完成，例如使用 Linux 的 Systemd 系统进行管理。

注意，如果之前已运行过这 3 个进程，则需要先停止它们，然后启动 kubelet 服务，这 3 个主进程将通过 kubelet 以容器的形式启动和运行。

接下来分别对 kube-apiserver 和 kube-controller-manager、kube-scheduler 的高可用部署进行说明。

### 1) kube-apiserver 的高可用部署

根据第 2 章的介绍，为 kube-apiserver 预先创建所有需要的 CA 证书和基本鉴权文件等内容，然后在每台服务器上创建其日志文件：

```
# touch /var/log/kube-apiserver.log
```

假设 kubelet 的启动参数指定--config=/etc/kubernetes/manifests，即 Static Pod 定义文件所在的目录，接下来就可以创建 kube-apiserver.yaml 配置文件用于启动 kube-apiserver 了。

```

kube-apiserver.yaml
apiVersion: v1
kind: Pod
metadata:
  name: kube-apiserver
spec:
  hostNetwork: true
  containers:
  - name: kube-apiserver
    image:
gcr.io/google_containers/kube-apiserver:9680e782e08a1a1c94c656190011bd02
    command:
    - /bin/sh
    - -c
    - /usr/local/bin/kube-apiserver --etcd-servers=http://127.0.0.1:2379
      --admission-control=NamespaceLifecycle,LimitRanger,SecurityContextDeny,ServiceAccount,ResourceQuota
      --service-cluster-ip-range=169.169.0.0/16 --v=2
      --allow-privileged=False 1>>/var/log/kube-apiserver.log 2>&1
    ports:
    - containerPort: 443
      hostPort: 443
      name: https
    - containerPort: 7080
      hostPort: 7080
      name: http
    - containerPort: 8080

```

```
    hostPort: 8080
    name: local
  volumeMounts:
  - mountPath: /srv/kubernetes
    name: srvkube
    readOnly: true
  - mountPath: /var/log/kube-apiserver.log
    name: logfile
  - mountPath: /etc/ssl
    name: etcssl
    readOnly: true
  - mountPath: /usr/share/ssl
    name: usrsharessl
    readOnly: true
  - mountPath: /var/ssl
    name: varssl
    readOnly: true
  - mountPath: /usr/ssl
    name: usrssl
    readOnly: true
  - mountPath: /usr/lib/ssl
    name: usrlibssl
    readOnly: true
  - mountPath: /usr/local/openssl
    name: usrlocalopenssl
    readOnly: true
  - mountPath: /etc/openssl
    name: etcopenssl
    readOnly: true
  - mountPath: /etc/pki/tls
    name: etcpkits
    readOnly: true
  volumes:
  - hostPath:
      path: /srv/kubernetes
      name: srvkube
  - hostPath:
      path: /var/log/kube-apiserver.log
      name: logfile
  - hostPath:
      path: /etc/ssl
      name: etcssl
  - hostPath:
      path: /usr/share/ssl
      name: usrsharessl
  - hostPath:
      path: /var/ssl
```

```

    name: varssl
- hostPath:
    path: /usr/ssl
    name: usrssl
- hostPath:
    path: /usr/lib/ssl
    name: usrlibssl
- hostPath:
    path: /usr/local/openssl
    name: usrlocalopenssl
- hostPath:
    path: /etc/openssl
    name: etcopenssl
- hostPath:
    path: /etc/pki/tls
    name: etcpkits

```

其中，

- ◎ kube-apiserver 需要使用 hostNetwork 模式，即直接使用宿主机网络，以使得客户端能够通过物理机访问其 API。
- ◎ 镜像的 tag 来源于 kubernetes 发布包中的 kube-apiserver.docker\_tag 文件：kubernetes/server/kubernetes-server-linux-amd64/server/bin/kube-apiserver.docker\_tag。
- ◎ --etcd-servers：指定 etcd 服务的 URL 地址。
- ◎ 再加上其他必要的启动参数，包括--admission-control、--service-cluster-ip-range、CA 证书相关配置等内容。
- ◎ 端口号的设置都配置了 hostPort，将容器内的端口号直接映射为宿主机的端口号。

将 kube-apiserver.yaml 文件复制到 kubelet 监控的/etc/kubernetes/manifests 目录下，kubelet 将会自动创建 yaml 文件中定义的 kube-apiserver 的 Pod。

接下来在另外两台服务器上重复该操作，使得每台服务器上都启动一个 kube-apiserver 的 Pod。

## 2) 为 kube-apiserver 配置负载均衡器

至此，我们启动了三个 kube-apiserver 实例，这三个 kube-apiserver 都可以正常工作，我们需要一个统一的、可靠的、允许部分 Master 节点故障的方式来访问它们，可以通过部署一个负载均衡器来实现。

在不同的平台下，负载均衡的实现方式不同：在一些公用云比如 GCE、AWS、阿里云上都有现成的实现方案；对于本地集群，我们可以选择硬件或者软件来实现负载均衡，比如 Kubernetes 社区推荐的方案 haproxy 和 keepalived 来实现，其中 haproxy 负责负载均衡，而 keepalived 负责对 haproxy 监控和进行高可用。

在完成 API Server 的负载均衡配置之后，对其访问还需要注意以下内容。

- ◎ 如果 Master 开启了安全认证机制，那么需要确保证书中包含负载均衡服务节点的 IP。
- ◎ 对于外部的访问，比如通过 kubectl 访问 API Server，那么需要配置为访问 API Server 对应的负载均衡器的 IP 地址。

### 3) kube-controller-manager 和 kube-scheduler 的高可用配置

不同于 API Server，Master 中另外两个核心组件 kube-controller-manager 和 kube-scheduler 会修改集群的状态信息，因此对于 kube-controller-manager 和 kube-scheduler 而言，高可用不仅意味着需要启动多个实例，还需要这多个实例能实现选举并选举出 leader，以保证同一时间只有一个实例可以对集群状态信息进行读写，避免出现同步问题和一致性问题。Kubernetes 对于这种选举机制的实现是采用租赁锁（lease-lock）来实现的，我们可以通过在 kube-controller-manager 和 kube-scheduler 的每个实例的启动参数中设置--leader-elect=true，来保证同一时间只会运行一个可修改集群信息的实例。

Scheduler 和 Controller Manager 高可用的具体实现方式如下。

首先在每个 Master 节点上创建相应的日志文件：

```
# touch /var/log/kube-scheduler.log
# touch /var/log/kube-controller-manager.log
```

然后创建 kube-controller-manager 和 kube-scheduler 的 Pod 定义文件：

```
kube-controller-manager.yaml:
apiVersion: v1
kind: Pod
metadata:
  name: kube-controller-manager
spec:
  hostNetwork: true
  containers:
  - name: kube-controller-manager
    image: gcr.io/google_containers/kube-controller-manager:
fda24638d51a48baa13c35337fcd4793
    command:
    - /bin/sh
    - -c
    - /usr/local/bin/kube-controller-manager --master=127.0.0.1:8080
      --v=2 --leader-elect=true 1>>/var/log/kube-controller-manager.log 2>&1
  livenessProbe:
    httpGet:
      path: /healthz
      port: 10252
    initialDelaySeconds: 15
```

```

    timeoutSeconds: 1
  volumeMounts:
  - mountPath: /srv/kubernetes
    name: srvkube
    readOnly: true
  - mountPath: /var/log/kube-controller-manager.log
    name: logfile
  - mountPath: /etc/ssl
    name: etcssl
    readOnly: true
  - mountPath: /usr/share/ssl
    name: usrsharessl
    readOnly: true
  - mountPath: /var/ssl
    name: varssl
    readOnly: true
  - mountPath: /usr/ssl
    name: usrssl
    readOnly: true
  - mountPath: /usr/lib/ssl
    name: usrllibssl
    readOnly: true
  - mountPath: /usr/local/openssl
    name: usrlocalopenssl
    readOnly: true
  - mountPath: /etc/openssl
    name: etcopenssl
    readOnly: true
  - mountPath: /etc/pki/tls
    name: etcpkitls
    readOnly: true
  volumes:
  - hostPath:
      path: /srv/kubernetes
      name: srvkube
  - hostPath:
      path: /var/log/kube-controller-manager.log
      name: logfile
  - hostPath:
      path: /etc/ssl
      name: etcssl
  - hostPath:
      path: /usr/share/ssl
      name: usrsharessl
  - hostPath:
      path: /var/ssl
      name: varssl

```

```
- hostPath:
  path: /usr/ssl
  name: usrssl
- hostPath:
  path: /usr/lib/ssl
  name: usrlibssl
- hostPath:
  path: /usr/local/openssl
  name: usrlocalopenssl
- hostPath:
  path: /etc/openssl
  name: etcopenssl
- hostPath:
  path: /etc/pki/tls
  name: etcpkits
```

其中，

- ◎ kube-controller-manager 需要使用 hostNetwork 模式，即直接使用宿主机网络。
- ◎ 镜像的 tag 来源于 Kubernetes 发布包中的 kube-controller-manager.docker\_tag 文件：  
kubernetes/server/kubernetes-server-linux-amd64/server/bin/kube-controller-manager.docker\_tag。
- ◎ --master: 指定 kube-apiserver 服务的 URL 地址。
- ◎ --leader-elect=true: 使用 leader 选举机制。

**kube-scheduler.yaml:**

```
apiVersion: v1
kind: Pod
metadata:
  name: kube-scheduler
spec:
  hostNetwork: true
  containers:
  - name: kube-scheduler
    image:
```

**gcr.io/google\_containers/kube-scheduler:34d0b8f8b31e27937327961528739bc9**

```
command:
- /bin/sh
- -c
- /usr/local/bin/kube-scheduler --master=127.0.0.1:8080 --v=2
```

**--leader-elect=true** 1>>/var/log/kube-scheduler.log 2>&1

```
livenessProbe:
  httpGet:
    path: /healthz
    port: 10251
  initialDelaySeconds: 15
  timeoutSeconds: 1
```



```

volumeMounts:
- mountPath: /var/log/kube-scheduler.log
  name: logfile
- mountPath: /var/run/secrets/kubernetes.io/serviceaccount
  name: default-token-s8ejd
  readOnly: true
volumes:
- hostPath:
    path: /var/log/kube-scheduler.log
    name: logfile

```

其中，

- ◎ kube-scheduler 需要使用 hostNetwork 模式，即直接使用宿主机网络。
- ◎ 镜像的 tag 来源于 kubernetes 发布包中的 kube-scheduler.docker\_tag 文件: kubernetes/server/kubernetes-server-linux-amd64/server/bin/kube-scheduler.docker\_tag。
- ◎ --master: 指定 kube-apiserver 服务的 URL 地址。
- ◎ --leader-elect=true: 使用 leader 选举机制。

将这两个 yaml 文件复制到 kubelet 监控的 /etc/kubernetes/manifests 目录下, kubelet 将会自动创建 yaml 文件中定义的 kube-controller-manager 和 kube-scheduler 的 Pod。

至此，我们完成了 Kubernetes Master 组件高可用的完整配置，配合 etcd 存储的高可用，整个 Kubernetes 集群的高可用已经全部完成。最后，只需要确认集群中所有访问 API Server 的地方都已经将访问地址修改为负载均衡的地址，就可以保证集群高可用的正常工作了。

### 3. Master 高可用架构的演进

在当前的版本中，kubelet 可以设置 “--api-servers” 启动参数来指定多个 kube-apiserver，但是当第 1 个 kube-apiserver 不可用之后，kubelet 无法连接到后面的 kube-apiserver，也就是说只有第 1 个 kube-apiserver 起作用。如果这个问题得到解决，则 kubelet 无须通过额外的负载均衡器就能连接到多个 API Server 了。

另外，除了 kubelet，其他核心组件 kube-controller-manager、kube-scheduler 和 kube-proxy 都需要配置 kube-apiserver，目前它们的启动参数 “--master” 仅支持配置一个 kube-apiserver，还无法支持多个 kube-apiserver 的配置。

Kubernetes 计划在后续的版本中支持多个 Master 的配置，实现不需要负载均衡器的 Master 高可用架构。

### 5.1.8 Kubernetes 集群监控

#### 1. 通过 cAdvisor 页面查看容器的运行状态

开源软件 cAdvisor（Container Advisor）是用于监控容器运行状态的利器之一（cAdvisor 项目的主页为 <https://github.com/google/cadvisor>），它被用于多个与 Docker 相关的开源项目中。

在 Kubernetes 系统中，cAdvisor 已被默认集成到了 kubelet 组件内，当 kubelet 服务启动时，它会自动启动 cAdvisor 服务，然后 cAdvisor 会实时采集所在节点的性能指标及在节点上运行的容器的性能指标。kubelet 的启动参数--cadvisor-port 可自定义 cAdvisor 对外提供服务的端口号，默认为 4194。

cAdvisor 提供了 Web 页面可供浏览器访问。例如 Kubernetes 集群中的一个 Node 的 IP 地址是 192.168.18.3，则在浏览器中输入网址 <http://192.168.18.3:4194> 来访问 cAdvisor 的监控页面。cAdvisor 的主页显示了主机的实时运行状态，包括 CPU 使用情况、内存使用情况、网络吞吐量及文件系统使用情况等信息。

图 5.7 展示了 cAdvisor 的几个性能监控页面。

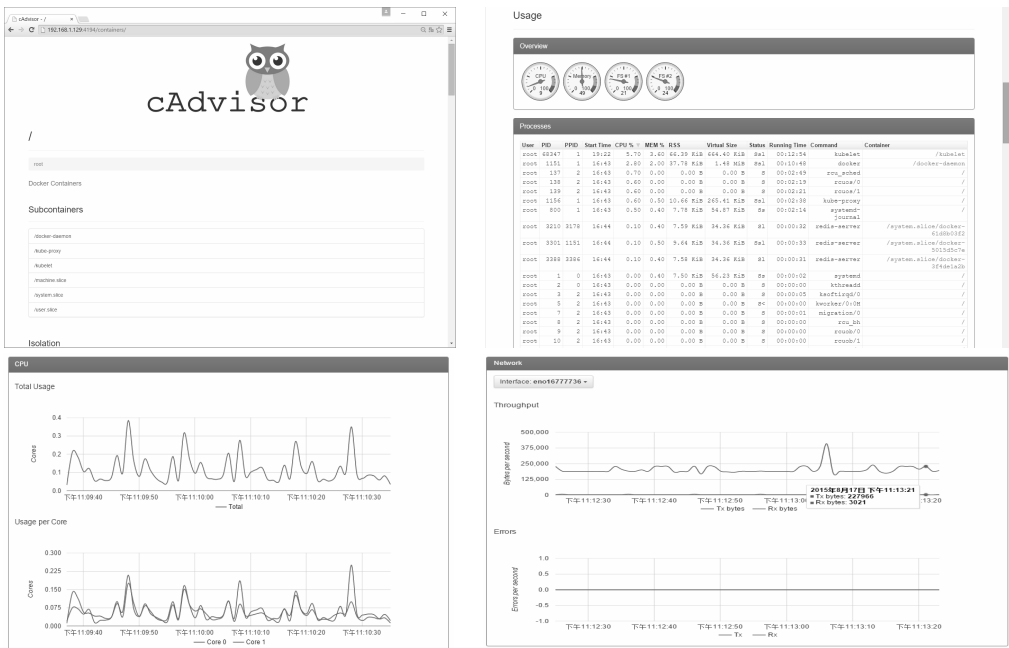


图 5.7 cAdvisor 的几个性能监控页面

通过 Docker Containers 链接可以查看容器列表及每个容器的性能数据，如图 5.8 所示。

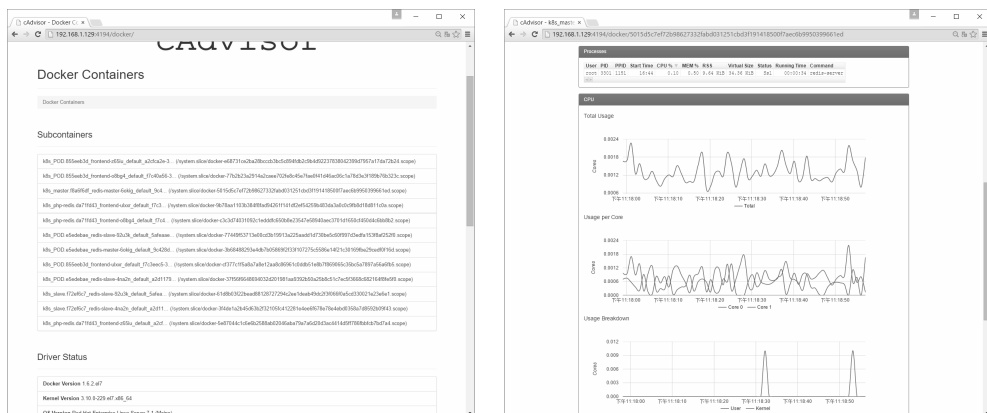


图 5.8 容器的性能监控页面

此外，cAdvisor 也提供了 REST API 供客户端远程调用，主要是为了定制开发，API 返回的数据格式为 JSON，可以采用如下 URL 来访问：

`http://<hostname>:<port>/api/<version>/<request>`

例如，通过 URL `http://192.168.18.3:4194/api/v1.3/machine` 可以获取主机的相关信息：

```
{
  "num_cores":2,
  "cpu_frequency_khz":2793544,
  "memory_capacity":1915408384,
  "machine_id":"0f6233d8256a4ec1a673640e04b8344a",
  "system_uuid":"564D188F-8E82-21C0-6E89-176E2C51EBB5",
  "boot_id":"a03d00d8-ca9c-4d74-a674-ebf5dfbc69d9",
  "filesystems":[
    {
      "device":"/dev/mapper/rhel-root",
      "capacity":18746441728
    },
    {
      "device":"/dev/sda1",
      "capacity":520794112
    }
  ],
  "disk_map":{
    "253:0":{
      "name":"dm-0",
      "major":253,
      "minor":0,
      "size":2147483648,
      "scheduler":"none"
    },
  },
}
```

```
        .....
    },
    "network_devices":[
        {
            "name":"eno16777736",
            "mac_address":"00:0c:29:51:eb:b5",
            "speed":1000,
            "mtu":1500
        }
    ],
    "topology":[
        {
            "node_id":0,
            "memory":2146947072,
            "cores":[
                {
                    "core_id":0,
                    "thread_ids":[
                        0
                    ],
                    "caches":null
                },
                .....
            ],
            "caches":[
                {
                    "size":6291456,
                    "type":"Unified",
                    "level":3
                }
            ]
        }
    ]
}
```

通过下面的 URL 则可以获取节点上最近 1min 的容器性能数据：<http://192.168.1.129:4194/api/v1.3/subcontainers/system.slice/docker-5015d5c7ef72b98627332fabd031251cbd3f191418500f7aec6b9950399661ed.scope>。

结果为：

```
[
  {
    "name":"/system.slice/docker-5015d5c7ef72b98627332fabd031251cbd3f191418500f7aec6b9950399661ed.scope",
    "aliases":[
      "k8s_master.f8a6f6df_Redis-master-6okig_default_9c428d4f-4167-11e5-afe7-000c
```

```

2921ba71_5dce2f85",
  "5015d5c7ef72b98627332fabd031251cbd3f191418500f7aec6b9950399661ed"
],
  "namespace": "docker",
  "spec": {
    "creation_time": "2015-08-17T08:44:27.401122502Z",
    "labels": {
      "io.kubernetes.pod.name": "default/Redis-master-6okig"
    },
    "has_cpu": true,
    "cpu": {
      "limit": 2,
      "max_limit": 0,
      "mask": "0-1"
    },
    "has_memory": true,
    "memory": {
      "limit": 18446744073709552000,
      "swap_limit": 18446744073709552000
    },
    "has_network": true,
    "has_filesystem": false,
    "has_diskio": true
  },
  "stats": [
    {
      "timestamp": "2015-08-18T00:54:26.167988505+08:00",
      "cpu": {
        "usage": {
          "total": 43121463207,
          "per_cpu_usage": [
            21578091763,
            21543371444
          ],
          "user": 410000000,
          "system": 13620000000
        },
        "load_average": 0
      },
      "diskio": {
        "io_service_bytes": [
          {
            "major": 253, "minor": 14,
            "stats": {
              "Async": 8036352, "Read": 8036352, "Sync": 0, "Total": 8036352, "Write": 0
            }
          }
        ]
      }
    }
  ]
}

```

```

    ],
    "io_serviced": [
      {
        "major": 8,
        "minor": 0,
        "stats": {
          "Async": 0,
          .....
        }
      ]
    },
    "memory": {
      "usage": 16748544,
      "working_set": 9297920,
      "container_data": {
        "pgfault": 882,
        "pgmajfault": 8
      },
      "hierarchical_data": {
        "pgfault": 882,
        "pgmajfault": 8
      }
    },
    "network": {
      "name": "",
      "rx_bytes": 0, "rx_packets": 0, "rx_errors": 0, "rx_dropped": 0, "tx_bytes": 0, "tx_packets": 0, "tx_errors": 0, "tx_dropped": 0
    },
    "task_stats": {
      "nr_sleeping": 0, "nr_running": 0, "nr_stopped": 0, "nr_uninterruptible": 0, "nr_io_wait": 0
    }
  },
  .....
]
}
]

```

容器的性能数据对于集群监控非常有用，系统管理员可以根据 cAdvisor 提供的数据进行分析 and 告警。不过，由于 cAdvisor 是在每台 Node 上运行的，只能采集本机的性能指标数据。对于大规模集群，Kubernetes 建议使用几个开源软件组成的集成解决方案来实现对整个集群的监控。

## 2. Heapster+Influxdb+Grafana 集群性能监控平台搭建

根据上节的说明，cAdvisor 集成在 kubelet 中，运行在每个 Node 上，所以一个 cAdvisor

仅能对一台 Node 进行监控。在大规模容器集群中，需要对所有 Node 和全部容器进行性能监控，Kubernetes 建议使用一套工具来实现集群性能数据的采集、存储和展示：Heapster、InfluxDB 和 Grafana。

- ◎ **Heapster**: 对集群中各 Node 上 cAdvisor 的数据采集汇聚的系统，通过访问每个 Node 上 kubelet 的 API，再通过 kubelet 调用 cAdvisor 的 API 来采集该节点上所有容器的性能数据。Heapster 对性能数据进行聚合，并将结果保存到后端存储系统中。Heapster 支持多种后端存储系统，包括 memory（保存在内存中）、InfluxDB、BigQuery、谷歌云平台提供的 Google Cloud Monitoring (<https://cloud.google.com/monitoring/>) 和 Google Cloud Logging (<https://cloud.google.com/logging/>) 等。Heapster 项目的主页为 <https://github.com/kubernetes/heapster>。
- ◎ **InfluxDB**: 是分布式时序数据库（每条记录都带有时间戳属性），主要用于实时数据采集、事件跟踪记录、存储时间图表、原始数据等。InfluxDB 提供了 REST API 用于数据的存储和查询。InfluxDB 的主页为 <http://influxdb.com>。
- ◎ **Grafana**: 通过 Dashboard 将 InfluxDB 中的时序数据展现成图表或曲线等形式，便于运维人员查看集群的运行状态。Grafana 的主页为 <http://grafana.org>。

基于 Heapster+InfluxDB+Grafana 的集群监控系统总体架构如图 5.9 所示。

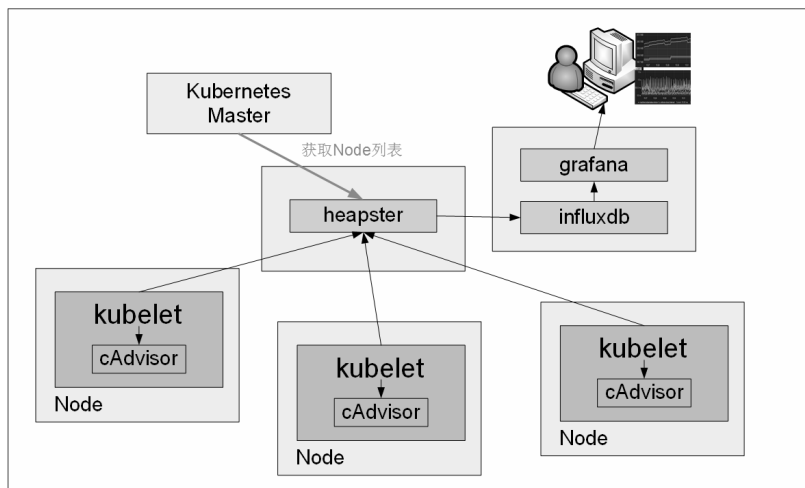


图 5.9 基于 Heapster+InfluxDB+Grafana 的集群监控系统总体架构

Heapster、InfluxDB 和 Grafana 均以 Pod 的形式启动和运行。由于 Heapster 需要与 Kubernetes Master 进行安全连接，所以需要设置 Master 的 CA 证书安全策略（参见第 2 章的说明）。

## 1) 部署 Heapster 容器

Heapster 的 Service 和 Deployment 定义如下：

### heapster.yaml

```
---
kind: Service
apiVersion: v1
metadata:
  name: heapster
  namespace: kube-system
  labels:
    kubernetes.io/cluster-service: "true"
    addonmanager.kubernetes.io/mode: Reconcile
    kubernetes.io/name: "Heapster"
spec:
  ports:
    - port: 80
      targetPort: 8082
  selector:
    k8s-app: heapster

---
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: heapster-v1.3.0
  namespace: kube-system
  labels:
    k8s-app: heapster
    kubernetes.io/cluster-service: "true"
    addonmanager.kubernetes.io/mode: Reconcile
    version: v1.3.0
spec:
  replicas: 1
  selector:
    matchLabels:
      k8s-app: heapster
      version: v1.3.0
  template:
    metadata:
      labels:
        k8s-app: heapster
        version: v1.3.0
      annotations:
        scheduler.alpha.kubernetes.io/critical-pod: ''
    spec:
```



```

containers:
- image: gcr.io/google_containers/heapster-amd64:v1.3.0
  name: heapster
  livenessProbe:
    httpGet:
      path: /healthz
      port: 8082
      scheme: HTTP
    initialDelaySeconds: 180
    timeoutSeconds: 5
  command:
    - /heapster
    - --source=kubernetes.summary_api:''
    - --sink=influxdb:http://monitoring-influxdb:8086
- image: gcr.io/google_containers/heapster-amd64:v1.3.0
  name: eventer
  command:
    - /eventer
    - --source=kubernetes:''
    - --sink=influxdb:http://monitoring-influxdb:8086
- image: gcr.io/google_containers/addon-resizer:1.7
  name: heapster-nanny
  resources:
    limits:
      cpu: 50m
      memory: 90Mi
    requests:
      cpu: 50m
      memory: 90Mi
  env:
    - name: MY_POD_NAME
      valueFrom:
        fieldRef:
          fieldPath: metadata.name
    - name: MY_POD_NAMESPACE
      valueFrom:
        fieldRef:
          fieldPath: metadata.namespace
  command:
    - /pod_nanny
    - --cpu=80m
    - --extra-cpu=0m
    - --memory=140Mi
    - --extra-memory=4Mi
    - --threshold=5
    - --deployment=heapster-v1.3.0
    - --container=heapster

```

```
- --poll-period=300000
- --estimator=exponential
- image: gcr.io/google_containers/addon-resizer:1.7
  name: eventer-nanny
  resources:
    limits:
      cpu: 50m
      memory: 90Mi
    requests:
      cpu: 50m
      memory: 90Mi
  env:
    - name: MY_POD_NAME
      valueFrom:
        fieldRef:
          fieldPath: metadata.name
    - name: MY_POD_NAMESPACE
      valueFrom:
        fieldRef:
          fieldPath: metadata.namespace
  command:
    - /pod_nanny
    - --cpu=100m
    - --extra-cpu=0m
    - --memory=140Mi
    - --extra-memory=500Ki
    - --threshold=5
    - --deployment=heapster-v1.3.0
    - --container=eventer
    - --poll-period=300000
    - --estimator=exponential
  tolerations:
    - key: "CriticalAddonsOnly"
      operator: "Exists"
```

Heapster 需要设置的启动参数如下。

#### (1) --source

配置采集来源，为 Master URL 地址。设置为空字符串表示使用内置服务 <https://kubernetes:443> 连接 API，也可以指定 IP 地址和端口号进行连接。

```
--source=kubernetes.summary_api:''
```

#### (2) --sink

配置后端存储系统，使用 InfluxDB 数据库：

```
--sink=InfluxDB:http://monitoring-InfluxDB:8086
```

注意, InfluxDB 数据库地址使用的是 InfluxDB 的 Service 名字, 这需要 DNS 服务正常工作, 如果没有配置 DNS 服务, 则也可以使用 Service 的 ClusterIP 地址。

### (3) --metric\_resolution

性能指标的精度, 60s 表示将过去 60s 的数据进行汇聚后再进行存储。

其他参数可以通过进入 Heapster 容器执行 `# heapster --help` 命令查看和设置。

使用 `kubectrl create` 命令创建 Heapster:

```
# kubectrl create -f heapster.yaml
deployment "heapster-v1.3.0" created
service "heapster" created
```

## 2) 部署 Influxdb 和 Grafana

Influxdb Service 定义如下:

### **influxdb-service.yaml**

```
apiVersion: v1
kind: Service
metadata:
  name: monitoring-influxdb
  namespace: kube-system
  labels:
    kubernetes.io/cluster-service: "true"
    addonmanager.kubernetes.io/mode: Reconcile
    kubernetes.io/name: "InfluxDB"
spec:
  type: NodePort
  ports:
    - name: http
      port: 8083
      targetPort: 8083
      nodePort: 8083
    - name: api
      port: 8086
      targetPort: 8086
  selector:
    k8s-app: influxGrafana
```

注意, 这里使用 `type=NodePort` 将 InfluxDB 的 Web 服务映射在宿主机 Node 的端口上, 以便我们使用浏览器对其进行访问。

Grafana Service 定义如下:

### **grafana-service.yaml**

```
apiVersion: v1
```

```
kind: Service
metadata:
  name: monitoring-grafana
  namespace: kube-system
  labels:
    kubernetes.io/cluster-service: "true"
    addonmanager.kubernetes.io/mode: Reconcile
    kubernetes.io/name: "Grafana"
spec:
  ports:
    - port: 80
      targetPort: 3000
  selector:
    k8s-app: influxGrafana
```

InfluxDB 和 Grafana 的 RC 定义如下：

#### **influxdb-grafana-rc.yaml**

```
apiVersion: v1
kind: ReplicationController
metadata:
  name: monitoring-influxdb-grafana-v4
  namespace: kube-system
  labels:
    k8s-app: influxGrafana
    version: v4
    kubernetes.io/cluster-service: "true"
    addonmanager.kubernetes.io/mode: Reconcile
spec:
  replicas: 1
  selector:
    k8s-app: influxGrafana
    version: v4
  template:
    metadata:
      labels:
        k8s-app: influxGrafana
        version: v4
        kubernetes.io/cluster-service: "true"
    spec:
      containers:
        - image: gcr.io/google_containers/heapster-influxdb-amd64:v1.1.1
          name: influxdb
          resources:
            # keep request = limit to keep this container in guaranteed class
            limits:
              cpu: 100m
```

```

        memory: 500Mi
      requests:
        cpu: 100m
        memory: 500Mi
    ports:
      - containerPort: 8083
      - containerPort: 8086
    volumeMounts:
      - name: influxdb-persistent-storage
        mountPath: /data
- image: gcr.io/google_containers/heapster-grafana-amd64:v4.0.2
  name: grafana
  env:
  resources:
    # keep request = limit to keep this container in guaranteed class
    limits:
      cpu: 100m
      memory: 100Mi
    requests:
      cpu: 100m
      memory: 100Mi
  env:
    # This variable is required to setup templates in Grafana.
    - name: INFLUXDB_SERVICE_URL
      value: http://monitoring-influxdb:8086
    # The following env variables are required to make Grafana accessible
via
    # the kubernetes api-server proxy. On production clusters, we
recommend
    # removing these env variables, setup auth for grafana, and expose
the grafana
    # service using a LoadBalancer or a public IP.
    - name: GF_AUTH_BASIC_ENABLED
      value: "false"
    - name: GF_AUTH_ANONYMOUS_ENABLED
      value: "true"
    - name: GF_AUTH_ANONYMOUS_ORG_ROLE
      value: Admin
    - name: GF_SERVER_ROOT_URL
      value:
/api/v1/proxy/namespaces/kube-system/services/monitoring-grafana/
    volumeMounts:
      - name: grafana-persistent-storage
        mountPath: /var
  volumes:
    - name: influxdb-persistent-storage
      emptyDir: {}

```

```
- name: grafana-persistent-storage
  emptyDir: {}
```

注意，Grafana 容器环境变量 `INFLUXDB_SERVICE_URL` 设置为 InfluxDB 服务的所在地址。由于 Grafana 与 InfluxDB 处于同一个 Pod 中，所以 Grafana 使用 `127.0.0.1` 或 `localhost` 也可以访问到 InfluxDB 服务。

另外，对于 InfluxDB 和 Grafana 容器的 CPU 和内存资源请求和资源限制的值，需要根据集群规模进行调整。

使用 `kubectl create` 命令创建 InfluxDB 和 Grafana：

```
# kubectl create -f influxdb-service.yaml
service "monitoring-influxdb" created
# kubectl create -f grafana-service.yaml
service "monitoring-grafana" created
# kubectl create -f influxdb-grafana-rc.yaml
replicationcontroller "monitoring-influxdb-grafana-v4" created
```

使用 `kubectl get pods --namespace=kube-system` 命令确认 Pod 成功启动：

```
# kubectl get pods --namespace=kube-system
```

NAME	READY	STATUS	RESTARTS	AGE
heapster-v1.3.0-1626276742-k6gps	4/4	Running	0	3m
monitoring-influxdb-grafana-v4-cmcbw	2/2	Running	0	5m
kube-dns-v11-rxf5k	4/4	Running	0	10m

查看 Heapster 的日志，确保 Heapster 成功在 InfluxDB 数据库中创建名为 `k8s` 的数据库：

```
# kubectl logs heapster-v1.1.0-1895667918-guis1 -c heapster
--namespace=kube-system
I0706 03:43:50.212241      1 heapster.go:72] /heapster
--source=kubernetes.summary_api:''
--sink=influxdb:http://monitoring-influxdb:8086
I0706 03:43:50.212290      1 heapster.go:73] Heapster version v1.3.0
I0706 03:43:50.212470      1 configs.go:61] Using Kubernetes client with master
"https://169.169.0.1:443" and version v1
I0706 03:43:50.212481      1 configs.go:62] Using kubelet port 10255
I0706 03:43:50.725878      1 influxdb.go:252] created influxdb sink with options:
host:monitoring-influxdb:8086 user:root db:k8s
I0706 03:43:50.725950      1 heapster.go:196] Starting with InfluxDB Sink
I0706 03:43:50.725972      1 heapster.go:196] Starting with Metric Sink
I0706 03:43:52.104438      1 heapster.go:106] Starting heapster on port 8082
```

### 3) 查询 InfluxDB 数据库中的数据

让我们先通过 InfluxDB 的管理页面查看数据。

由于设置 InfluxDB 服务会暴露到物理 Node 节点上，所以我们可以通过任一 Node 的 8083 端口访问 InfluxDB 数据库提供的管理页面，如图 5.10 所示。通过右上角的齿轮按钮可以修改连

接属性（用于 InfluxDB service 设置为非默认端口号时）。单击右上角的 Database 下拉列表可以选择数据库，Heapster 创建的数据库名为 k8s。

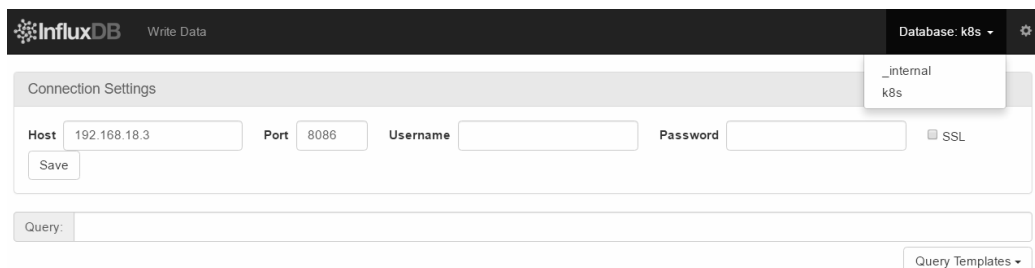


图 5.10 InfluxDB 管理页面

在 Query 输入框中输入“SHOW MEASUREMENTS”，即可查看所有的 measurements（序列表）。图 5.11 显示了部分 measurements。

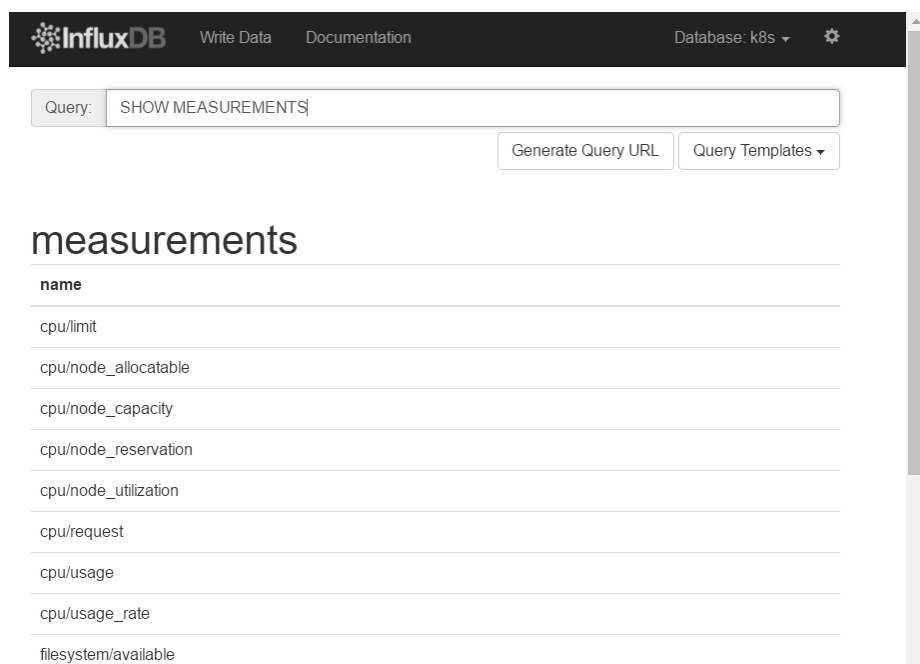


图 5.11 SHOW MEASUREMENTS 的部分结果页面

Heapster 采集的全部 metric（性能指标）如表 5.11 所示。

表 5.11 Heapster 采集的 metric

metric 名称	说 明
cpu/limit	CPU hard limit，单位为 ms
cpu/node_reservation	Node 保留的 CPU Share
cpu/node_utilization	Node 的 CPU 使用时间
cpu/request	CPU request，单位为 ms
cpu/usage	全部 Core 的 CPU 累计使用时间
cpu/usage_rate	全部 Core 的 CPU 累计使用率，单位为 ms
filesystem/usage	文件系统已用的空间，单位为字节
filesystem/limit	文件系统总空间限制，单位为字节
filesystem/available	文件系统可用的空间，单位为字节
memory/limit	Memory hard limit，单位为字节
memory/major_page_faults	major page faults 数量
memory/major_page_faults_rate	每秒的 major page faults 数量
memory/node_reservation	Node 保留的内存 Share
memory/node_utilization	Node 的内存使用值
memory/page_faults	page faults 数量
memory/page_faults_rate	每秒的 page faults 数量
memory/request	Memory request，单位为字节
memory/usage	总内存使用量
memory/working_set	总的 Working set usage，Working set 是指不会被 kernel 移除的内存
network/rx	累计接收的网络流量字节数
network/rx_errors	累计接收的网络流量错误数
network/rx_errors_rate	每秒接收的网络流量错误数
network/rx_rate	每秒接收的网络流量字节数
network/tx	累计发送的网络流量字节数
network/tx_errors	累计发送的网络流量错误数
network/tx_errors_rate	每秒发送的网络流量错误数
network/tx_rate	每秒发送的网络流量字节数
uptime	容器启动总时长

每个 metric 可以看作一张数据库表，表中每条记录由一组 label 组成，可以看作字段，如表 5.12 所示。



表 5.12 metric 的各 label

Label 名称	说 明
pod_id	系统生成的 Pod 唯一名称
pod_name	用户指定的 Pod 名称
pod_namespace	Pod 所属的 namespace
container_base_image	容器的镜像名称
container_name	用户指定的容器名称
host_id	用户指定的 Node 主机名
hostname	容器运行所在主机名
labels	逗号分隔的 Label 列表
namespace_id	Pod 所属的 namespace 的 UID
resource_id	资源 ID

可以使用标准 SQL SELECT 语句对每个 metric 进行查询，例如查询 CPU 的使用时间：

```
select * from "cpu/usage" limit 10
```

结果如图 5.12 所示。

time	container_base_image	container_name	host_id	hostname	labels	namespace_id	namespace_name	nodename	pod_id	pod_name	pod_namespace	type	value
2016-08-06T21:32:00Z	"gcr.io/google_containers/heapster-v1.1.0"	"eveiter"	"k8s-node-1"	"k8s-node-1"	"k8s-app/heapster-pod-template-hash:1895667918.version.v1.1.0"	"7950e852-4a50-11e6-ba0c-000c296c2102"	"kube-system"	"k8s-node-1"	"31ef1604-5c1d-11e6-bca7-000c296c2102"	"heapster-v1.1.0-1895667918-bca7-000c296c2102"	"kube-system"	"pod_container"	43040357
2016-08-06T21:32:00Z	"gcr.io/google_containers/skydns:2015-10-13-8c728c"	"skydns"	"k8s-node-1"	"k8s-node-1"	"k8s-app/kube-dns.kubernetes.io/cluster-service.true.version.v11"	"7950e852-4a50-11e6-ba0c-000c296c2102"	"kube-system"	"k8s-node-1"	"ca76b813-5b02-11e6-bca7-000c296c2102"	"kube-dns-v11-q2hc"	"kube-system"	"pod_container"	234138702161
2016-08-06T21:32:00Z	"gcr.io/google_containers/evchealth:1.0"	"healthz"	"k8s-node-1"	"k8s-node-1"	"k8s-app/kube-dns.kubernetes.io/cluster-service.true.version.v11"	"7950e852-4a50-11e6-ba0c-000c296c2102"	"kube-system"	"k8s-node-1"	"ca76b813-5b02-11e6-bca7-000c296c2102"	"kube-dns-v11-q2hc"	"kube-system"	"pod_container"	263245016079
2016-08-06T21:32:00Z	"gcr.io/google_containers/addon-resizer:1.3"	"heapster-nanny"	"k8s-node-1"	"k8s-node-1"	"k8s-app/heapster-pod-template-hash:1895667918.version.v1.1.0"	"7950e852-4a50-11e6-ba0c-000c296c2102"	"kube-system"	"k8s-node-1"	"31ef1604-5c1d-11e6-bca7-000c296c2102"	"heapster-v1.1.0-1895667918-bca7-000c296c2102"	"kube-system"	"pod_container"	27816092
2016-08-06T21:32:00Z			"k8s-node-1"	"k8s-node-1"				"k8s-node-1"				"node"	48032643618721
2016-08-06T21:32:00Z	"gcr.io/google_containers/heapster-v1.1.0"	"heapster"	"k8s-node-1"	"k8s-node-1"	"k8s-app/heapster-pod-template-hash:1895667918.version.v1.1.0"	"7950e852-4a50-11e6-ba0c-000c296c2102"	"kube-system"	"k8s-node-1"	"31ef1604-5c1d-11e6-bca7-000c296c2102"	"heapster-v1.1.0-1895667918-bca7-000c296c2102"	"kube-system"	"pod_container"	41665260
2016-08-06T21:32:00Z	"gcr.io/google_containers/kube2sky-amd64:1.15"	"kube2sky"	"k8s-node-1"	"k8s-node-1"	"k8s-app/kube-dns.kubernetes.io/cluster-service.true.version.v11"	"7950e852-4a50-11e6-ba0c-000c296c2102"	"kube-system"	"k8s-node-1"	"ca76b813-5b02-11e6-bca7-000c296c2102"	"kube-dns-v11-q2hc"	"kube-system"	"pod_container"	17674965052
2016-08-06T21:32:00Z	"gcr.io/google_containers/addon-resizer:1.3"	"eveiter-nanny"	"k8s-node-1"	"k8s-node-1"	"k8s-app/heapster-pod-template-hash:1895667918.version.v1.1.0"	"7950e852-4a50-11e6-ba0c-000c296c2102"	"kube-system"	"k8s-node-1"	"31ef1604-5c1d-11e6-bca7-000c296c2102"	"heapster-v1.1.0-1895667918-bca7-000c296c2102"	"kube-system"	"pod_container"	41565306
2016-08-06T21:32:00Z	"gcr.io/google_containers/heapster-influxdb:v0.5"	"influxdb"	"k8s-node-1"	"k8s-node-1"	"k8s-app/influx-grafana.kubernetes.io/cluster-service.true.version.v3"	"7950e852-4a50-11e6-ba0c-000c296c2102"	"kube-system"	"k8s-node-1"	"1d6ecec6-5c1d-11e6-bca7-000c296c2102"	"monitoring-influxdb-grafana-v3-bca7-000c296c2102"	"kube-system"	"pod_container"	427297277

图 5.12 查询 cpu/usage 的结果页面

#### 4) Grafana 页面查看和操作

访问 Grafana 服务时需要通过 Master 代理模式进行，URL 地址为 `http://192.168.18.3:8080/api/v1/proxy/namespaces/kube-system/services/monitoring-grafana/`。

在 Grafana 主页可以查看监控数据的图表展示画面。如图 5.13 所示为 Cluster 集群的整体信息，以折线图的形式展示了集群范围内各 Node 的 CPU 使用率、内存使用情况等信息。



图 5.13 Grafana Cluster 监控页面

图 5.14 显示的是所有 Pod 的信息，以折线图的形式展示了集群范围内各 Pod 的 CPU 使用率、内存使用情况、网络流量、文件系统使用情况等信息。



图 5.14 Grafana Pod 监控页面

Grafana 页面上的每个图表都可以进行编辑，在标题上单击鼠标，点击“Edit”进入编辑页面，可以对每个 metric 进行个性化设置，例如查询的表名、字段名、汇总计算等，如图 5.15 所示。

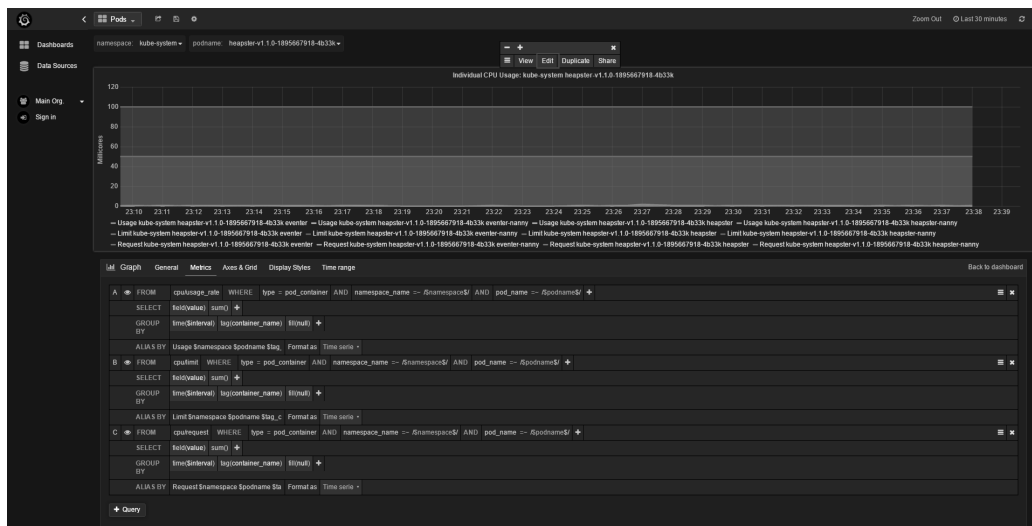


图 5.15 编辑页面

至此，基于 Heapster+InfluxDB+Grafana 的 Kubernetes 集群监控系统就搭建完成了。

### 5.1.9 集群统一日志管理

在 Kubernetes 集群环境中，一个完整的应用或服务都会涉及为数众多的组件运行，各组件所在的 Node 及实例数量都是可变的。日志子系统如果不做集中化管理，则会给系统的运维支撑造成很大的困难，因此有必要在集群层面对日志进行统一的收集和检索等工作。

容器中输出到控制台的日志，都会以 \*.json.log 的命名方式保存在 /var/lib/docker/containers/ 目录下，这样就给了我们进行日志采集和后续处理的基础。

Kubernetes 推荐采用 Fluentd+Elasticsearch+Kibana 完成对系统和容器日志的采集、查询和展现工作。

在部署统一日志管理系统之前，需要以下两个前提条件。

- ◎ API Server 正确配置了 CA 证书。
- ◎ DNS 服务启动运行。

### 1. 系统部署架构

系统的逻辑架构如图 5.16 所示。

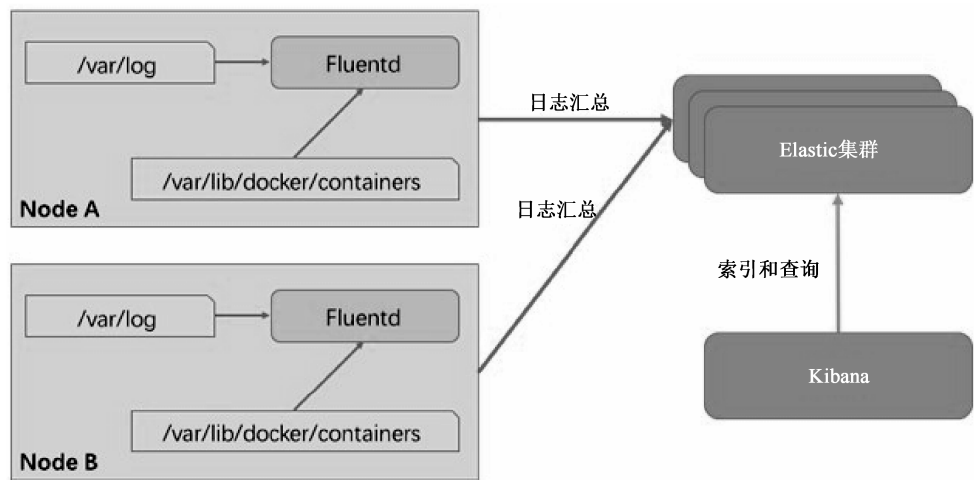


图 5.16 Fluentd+Elasticsearch+Kibana 系统的逻辑架构图

在各 Node 上运行一个 Fluentd 容器，采集本节点 /var/log 和 /var/lib/docker/containers 两个目录下的日志进程，然后汇总到 Elasticsearch 集群，最终通过 Kibana 完成和用户的交互工作。

这里有一个特殊的需求，Fluentd 必须在每个 Node 上运行一份，为了满足这一需要，我们有以下几种不同的方式来部署 Fluentd。

- ◎ 直接在 Node 主机上部署 Fluentd。
- ◎ 利用 kubelet 的 --config 参数，为每个 Node 加载 Fluentd Pod。
- ◎ 利用 DaemonSet 来让 Fluentd Pod 在每个 Node 上运行。

目前官方推荐的包括 Fluentd、Logstash 等日志或者监控类的 Pod 的运行方式就是 DaemonSet 方式，因此本节我们也以这一方式进行配置。

### 2. 创建 Elasticsearch RC 和 Service

Elasticsearch 的 RC 和 Service 定义如下：

```
elasticsearch-rc-svc.yml
---
apiVersion: v1
kind: ReplicationController
metadata:
```

```

name: elasticsearch-logging-v1
namespace: kube-system
labels:
  k8s-app: elasticsearch-logging
  version: v1
  kubernetes.io/cluster-service: "true"
spec:
  replicas: 2
  selector:
    k8s-app: elasticsearch-logging
    version: v1
  template:
    metadata:
      labels:
        k8s-app: elasticsearch-logging
        version: v1
        kubernetes.io/cluster-service: "true"
    spec:
      containers:
        - image: gcr.io/google_containers/elasticsearch:1.8
          name: elasticsearch-logging
          resources:
            # keep request = limit to keep this container in guaranteed class
            limits:
              cpu: 100m
            requests:
              cpu: 100m
          ports:
            - containerPort: 9200
              name: db
              protocol: TCP
            - containerPort: 9300
              name: transport
              protocol: TCP
          volumeMounts:
            - name: es-persistent-storage
              mountPath: /data
          volumes:
            - name: es-persistent-storage
              emptyDir: {}
      ---
    apiVersion: v1
    kind: Service
    metadata:
      name: elasticsearch-logging
      namespace: kube-system
      labels:

```

```
k8s-app: elasticsearch-logging
kubernetes.io/cluster-service: "true"
kubernetes.io/name: "Elasticsearch"
spec:
  ports:
  - port: 9200
    protocol: TCP
    targetPort: db
  selector:
    k8s-app: elasticsearch-logging
```

执行 `kubectl create -f elastic-search.yml` 命令完成创建。

命令成功执行后，首先验证 Pod 的运行情况。通过 `kubectl get pods --namespaces=kube-system` 获取运行中的 Pod：

```
# kubectl get pods --namespaces=kube-system
NAMESPACE      NAME                                READY   STATUS    RESTARTS   AGE
kube-system     elasticsearch-logging-v1-59qvp     1/1     Running   0          18h
kube-system     elasticsearch-logging-v1-xnv14     1/1     Running   0          18h
```

接下来通过 Elasticsearch 的页面验证其功能。

执行 `# kubectl cluster-info` 命令获取 Elasticsearch 服务的地址：

```
# kubectl cluster-info
Elasticsearch is running at
http://192.168.18.3:8080/api/v1/proxy/namespaces/kube-system/services/elasticsearch-logging
```

接下来使用 `# kubectl proxy` 命令对 `apiserver` 进行代理，成功执行后输出如下：

```
# kubectl proxy
Starting to serve on 127.0.0.1:8001
```

这样我们就可以在浏览器上访问 URL 地址 `http://192.168.18.3:8001/api/v1/proxy/namespaces/kube-system/services/elasticsearch-logging`，来验证 Elasticsearch 的运行情况了，返回的内容是一个 JSON 文档：

```
{
  "status": 200,
  "name": "Elasticsearch",
  "cluster_name": "kubernetes-logging",
  "version": {
    "number": "1.5.2",
    "build_hash": "62ff9868b4c8a0c45860bebb259e21980778ab1c",
    "build_timestamp": "2015-04-27T09:21:06Z",
    "build_snapshot": false,
    "lucene_version": "4.10.4"
  },
}
```

```
"tagline": "You Know, for Search"
}
```

### 3. 在每个 Node 上启动 Fluentd

**Fluentd 的 DaemonSet 定义如下：**

```
fluentd-ds.yml
---
apiVersion: extensions/v1beta1
kind: DaemonSet
metadata:
  name: fluentd-cloud-logging
  namespace: kube-system
  labels:
    k8s-app: fluentd-cloud-logging
spec:
  template:
    metadata:
      namespace: kube-system
      labels:
        k8s-app: fluentd-cloud-logging
    spec:
      containers:
        - name: fluentd-cloud-logging
          image: gcr.io/google_containers/fluentd-elasticsearch:1.17
          resources:
            limits:
              cpu: 100m
              memory: 200Mi
          env:
            - name: FLUENTD_ARGS
              value: -q
          volumeMounts:
            - name: varlog
              mountPath: /var/log
              readOnly: false
            - name: containers
              mountPath: /var/lib/docker/containers
              readOnly: false
          volumes:
            - name: containers
              hostPath:
                path: /var/lib/docker/containers
            - name: varlog
              hostPath:
                path: /var/log
```

通过 `kubectl create` 命令创建 `Fluentd` 容器：

```
# kubectl create -f fluentd-ds.yml
```

查看创建的结果：

```
# kubectl get daemonset
```

NAME	DESIRED	CURRENT	NODE-SELECTOR	AGE
fluentd-cloud-logging	3	3	<none>	1h

```
# kubectl get pods
```

NAMESPACE	NAME	READY	STATUS	RESTARTS	AGE
	fluentd-cloud-logging-7tw9z	1/1	Running	0	18h
	fluentd-cloud-logging-aqdn1	1/1	Running	0	18h
	fluentd-cloud-logging-o4usx	1/1	Running	0	18h

结果显示 `Fluentd DaemonSet` 正常运行，启动 3 个 `Pod`，与集群中的 `Node` 数量一致。

接下来，使用 `# kubectl logs fluentd-cloud-logging-7tw9z` 命令查看 `Pod` 的日志，在 `Elasticsearch` 正常工作的情况下，我们会看到类似下面这样的日志内容：

```
# kubectl logs fluentd-cloud-logging-7tw9z
Connection opened to Elasticsearch cluster =>
{:host=>"elasticsearch-logging", :port=>9200, :scheme=>"http"}
```

说明 `Fluentd` 与 `Elasticsearch` 已经正确建立了连接。

## 4. 运行 Kibana

至此我们已经运行了 `Elasticsearch` 和 `Fluentd`，数据的采集和汇聚过程已经完成，接下来使用 `Kibana` 来展示和操作数据。

**Kibana 的 RC 和 Service 定义如下：**

```
kibana-rc-svc.yml
```

```
---
```

```
apiVersion: v1
kind: ReplicationController
metadata:
  name: kibana-logging-v1
  namespace: kube-system
  labels:
    k8s-app: kibana-logging
    version: v1
    kubernetes.io/cluster-service: "true"
spec:
  replicas: 1
  selector:
    k8s-app: kibana-logging
```



```

    version: v1
  template:
    metadata:
      labels:
        k8s-app: kibana-logging
        version: v1
        kubernetes.io/cluster-service: "true"
    spec:
      containers:
      - name: kibana-logging
        image: gcr.io/google_containers/kibana:1.3
        resources:
          # keep request = limit to keep this container in guaranteed class
          limits:
            cpu: 100m
          requests:
            cpu: 100m
        env:
          - name: "ELASTICSEARCH_URL"
            value: "http://elasticsearch-logging:9200"
        ports:
          - containerPort: 5601
            name: ui
            protocol: TCP
---
apiVersion: v1
kind: Service
metadata:
  name: kibana-logging
  namespace: kube-system
  labels:
    k8s-app: kibana-logging
    kubernetes.io/cluster-service: "true"
    kubernetes.io/name: "Kibana"
spec:
  ports:
    - port: 5601
      protocol: TCP
      targetPort: ui
  selector:
    k8s-app: kibana-logging

```

通过 `kubectl create -f kibana-rc-svc.yml` 命令创建 Kibana 的 RC 和 Service:

```

# kubectl create -f kibana-rc-svc.yml
replicationcontroller "kibana-logging-v1" created
service "kibana-logging" created

```

查看 Kibana 的运行情况：

```
# kubectl get pods
NAMESPACE   NAME                                READY   STATUS    RESTARTS   AGE
default     kibana-logging-v1-olagk            1/1     Running   0           1h

# kubectl get svc
NAME                CLUSTER-IP      EXTERNAL-IP   PORT(S)    AGE
kibana-logging      169.169.195.177 <none>        5601/TCP    1h

# kubectl get rc
NAME                DESIRED   CURRENT   AGE
kibana-logging-v1   1         1         1h
```

结果表明运行均已成功。通过 `kubectl cluster-info` 命令获取 Kibana 服务的 URL 地址：

```
# kubectl cluster-info
Kibana is running at http://127.0.0.1:8080/api/v1/proxy/namespaces/kube-system/
services/kibana-logging
```

同样通过 `kubectl proxy` 命令启动代理，在出现 `Starting to serve on 127.0.0.1:8001` 字样之后，用浏览器访问 URL 地址即可访问 Kibana 页面 `http://192.168.18.3:8001/api/v1/proxy/namespaces/kube-system/services/kibana-logging`。

第 1 次进入页面需要进行一些设置，如图 5.17 所示，选择所需选项后单击 `create`。

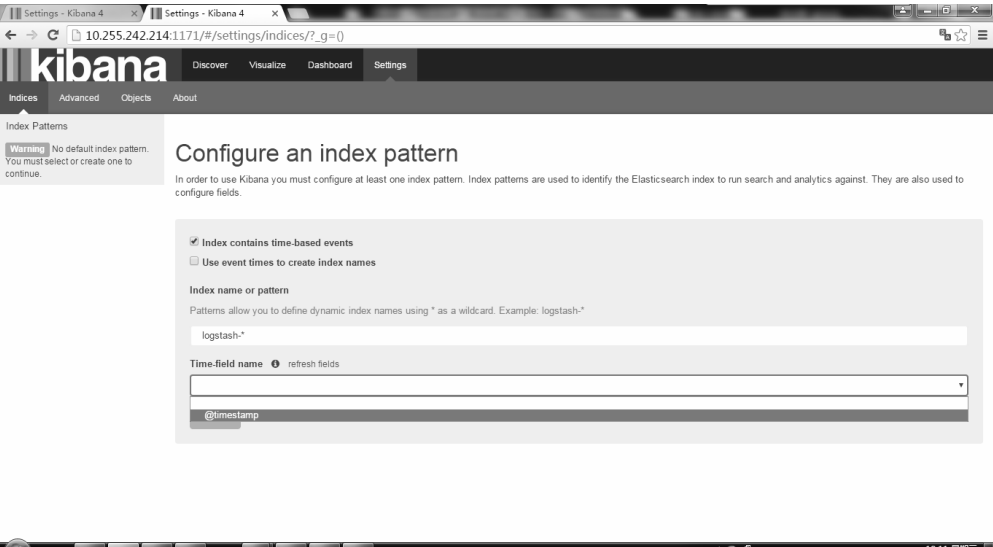


图 5.17 Kibana 创建索引页面

然后单击 `discover`，就可以正常查询日志了，如图 5.18 所示。

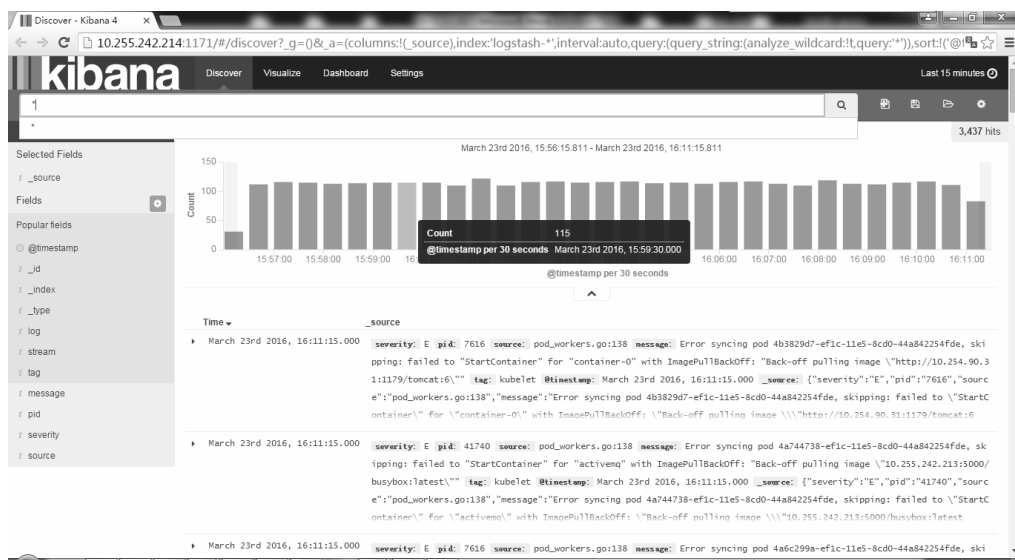


图 5.18 Kibana 查询日志页面

在搜索栏输入“error”关键字，可以搜索出从某些 Node 上找到的日志记录，如图 5.19 所示。

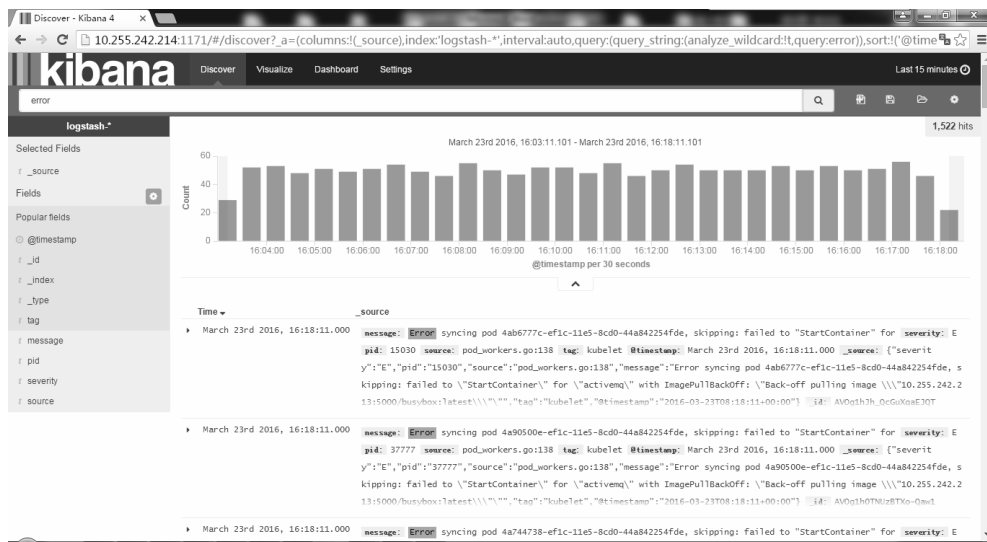


图 5.19 Kibana 日志关键字搜索页面

同时，通过左边菜单中 Fields 相关的内容对查询的内容进行限定，如图 5.20 所示。

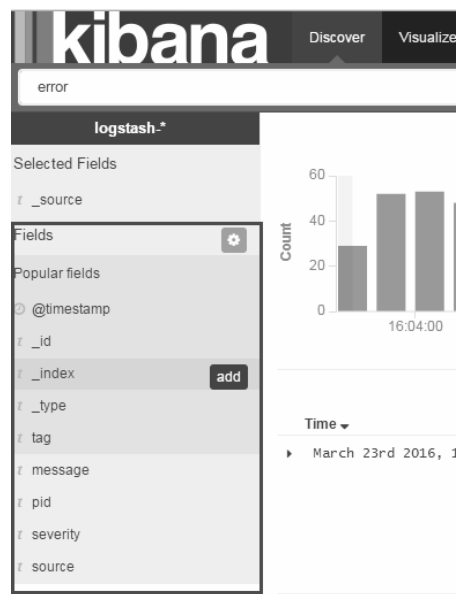


图 5.20 对查询的内容进行限定

至此，Kubernetes 集群范围内的统一日志收集和查询系统就搭建完成了。

### 5.1.10 Kubernetes 审计日志（Audit Log）

Kubernetes 从 v1.4 版本开始，为了加强对 Kubernetes 集群操作的安全监管，开始引入审计日志，以记录包括“什么时候？谁进行了什么操作？由哪个模块产生？发生了什么事？在哪里观察到？将引起什么结果？”等内容，便于系统管理员对集群中发生的各种事件进行追溯。目前，审计日志存在于 kube-apiserver 中，记录所有发往 API Server 的请求。每个审计日志包括请求和应答两条记录。

请求记录包括如下字段。

- ◎ 日志 ID，请求和应答具有相同的 ID。
- ◎ 请求的来源 IP 地址。
- ◎ 请求中的 HTTP 方法名。
- ◎ 调用方的原用户名。
- ◎ 调用方的原用户组信息。
- ◎ 操作用户名。

- ◎ 操作用户所属的组。
- ◎ 请求的 namespace。
- ◎ 请求的 URI 地址。

应答记录包括如下字段。

- ◎ 日志 ID，与请求日志具有相同的 ID。
- ◎ 返回码。

下例显示查询 Node k8s-node-1 状态请求的审计日志：

```
2017-04-23T13:50:57.603582097+08:00 AUDIT:
id="4b6cdaad-29e6-4b1d-b907-8e606a9477f6" ip="192.168.18.3" method="PATCH"
user="192.168.18.3" groups="\system:authenticated\" as="<self>"
asgroups="<lookup>" namespace="<none>" uri="/api/v1/nodes/k8s-node-1/status"
2017-04-23T13:50:57.618074625+08:00 AUDIT:
id="4b6cdaad-29e6-4b1d-b907-8e606a9477f6" response="200"
```

审计日志的相关配置参数都在 kube-apiserver 的启动参数中。

- ◎ `--audit-log-maxage`：审计日志文件保留的最长天数。
- ◎ `--audit-log-maxbackup`：审计日志文件的个数。
- ◎ `--audit-log-maxsize`：审计日志文件的单个大小限制，单位为 MB，默认为 100MB。
- ◎ `--audit-log-path`：审计日志文件的全路径。

审计日志文件以 `--audit-log-maxsize` 设置的大小为单位，写满后 kube-apiserver 将以时间戳重命名原文件，然后继续写入 `--audit-log-path` 指定的审计日志文件。`--audit-log-maxbackup` 和 `--audit-log-maxage` 参数则用于 kube-apiserver 自动删除旧的审计日志文件。

当前版本的审计日志提供的是对 API Server 请求的基本记录，更多的内容仍然在不断完善中。

### 5.1.11 使用 Web UI（Dashboard）管理集群

Kubernetes 的 Web UI 网页管理工具 `kubernetes-dashboard`，可提供部署应用、资源对象管理、容器日志查询、系统监控等常用的集群管理功能。为了在页面上显示系统资源的使用情况，要求部署 Heapster，详见前面章节的说明。

部署 `kubernetes-dashboard` 的 yaml 配置文件可通过 <https://rawgit.com/kubernetes/dashboard/master/src/deploy/kubernetes-dashboard.yaml> 页面下载。

配置文件 `kubernetes-dashboard.yaml` 的内容如下，包括 Deployment 和 Service 的定义：

## Kubernetes 权威指南：从 Docker 到 Kubernetes 实践全接触（纪念版）

```
kind: Deployment
apiVersion: extensions/v1beta1
metadata:
  labels:
    app: kubernetes-dashboard
    name: kubernetes-dashboard
    namespace: kube-system
spec:
  replicas: 1
  revisionHistoryLimit: 10
  selector:
    matchLabels:
      app: kubernetes-dashboard
  template:
    metadata:
      labels:
        app: kubernetes-dashboard
      # Comment the following annotation if Dashboard must not be deployed on master
      annotations:
        scheduler.alpha.kubernetes.io/tolerations: |
          [
            {
              "key": "dedicated",
              "operator": "Equal",
              "value": "master",
              "effect": "NoSchedule"
            }
          ]
    spec:
      containers:
        - name: kubernetes-dashboard
          image: gcr.io/google_containers/kubernetes-dashboard-amd64:v1.6.0
          imagePullPolicy: Always
          ports:
            - containerPort: 9090
              protocol: TCP
          args:
            # Uncomment the following line to manually specify Kubernetes API server
            Host
            # If not specified, Dashboard will attempt to auto discover the API server
            and connect
            # to it. Uncomment only if the default does not work.
            # - --apiserver-host=http://my-address:port
      livenessProbe:
        httpGet:
          path: /
          port: 9090
```

```

    initialDelaySeconds: 30
    timeoutSeconds: 30
---
kind: Service
apiVersion: v1
metadata:
  labels:
    app: kubernetes-dashboard
  name: kubernetes-dashboard
  namespace: kube-system
spec:
  type: NodePort
  ports:
    - port: 80
      targetPort: 9090
      nodePort: 9090
  selector:
    app: kubernetes-dashboard

```

这里，Service 设置了 NodePort 映射到物理机的端口号，用于客户端浏览器访问。

使用 `kubectl create` 命令进行部署：

```

# kubectl create -f kubernetes-dashboard.yaml
deployment "kubernetes-dashboard" created
service "kubernetes-dashboard" created

```

打开浏览器，输入某 Node 的 IP 和 9090 端口号，例如 `http://192.168.18.3:9090`，就能访问 dashboard 的页面了，如图 5.21 所示。

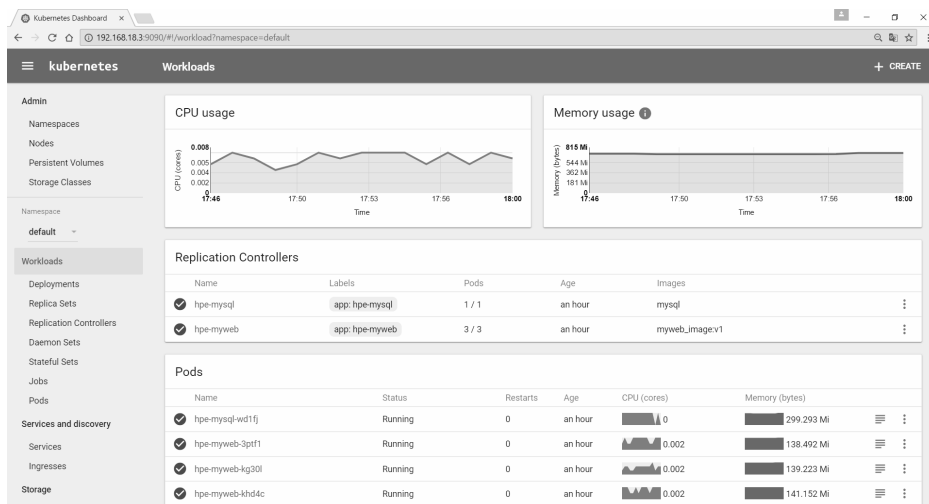


图 5.21 dashboard 页面

主页会默认显示 namespace 为 default 创建的 RC、Pod 列表，并显示各 Pod 的 CPU 和内存的性能指标。

单击右上角的“+ CREATE”按钮，将跳转到部署应用的页面。在这个页面上可以通过设置相关参数或者直接通过 yaml 或 json 文件进行应用的部署。如图 5.22 所示。

Create an app

+ CREATE

Deploy a Containerized App

☒ Specify app details below

☐ Upload a YAML or JSON file

App name \*

0 / 24

Container image \*

Number of pods \*

1

Service \*

None

SHOW ADVANCED OPTIONS

DEPLOY

CANCEL

To learn more, take the Dashboard Tour

An 'app' label with this value will be added to the Deployment and Service that get deployed. Learn more

Enter the URL of a public image on any registry, or a private image hosted on Docker Hub or Google Container Registry. Learn more

A Deployment will be created to maintain the desired number of pods across your cluster. Learn more

Optionally, an internal or external Service can be defined to map an incoming Port to a target Port seen by the container. Learn more

图 5.22 部署应用的页面

通过左侧的菜单，可以查看 Admin、Workloads、Service、Storage、Config 等各类资源对象的列表和详细信息。

例如，查看 Service 列表，如图 5.23 所示。

≡kubernetes

Services and discovery > Services

Pods

Services and discovery

Services

Ingresses

Storage

Persistent Volume Claims

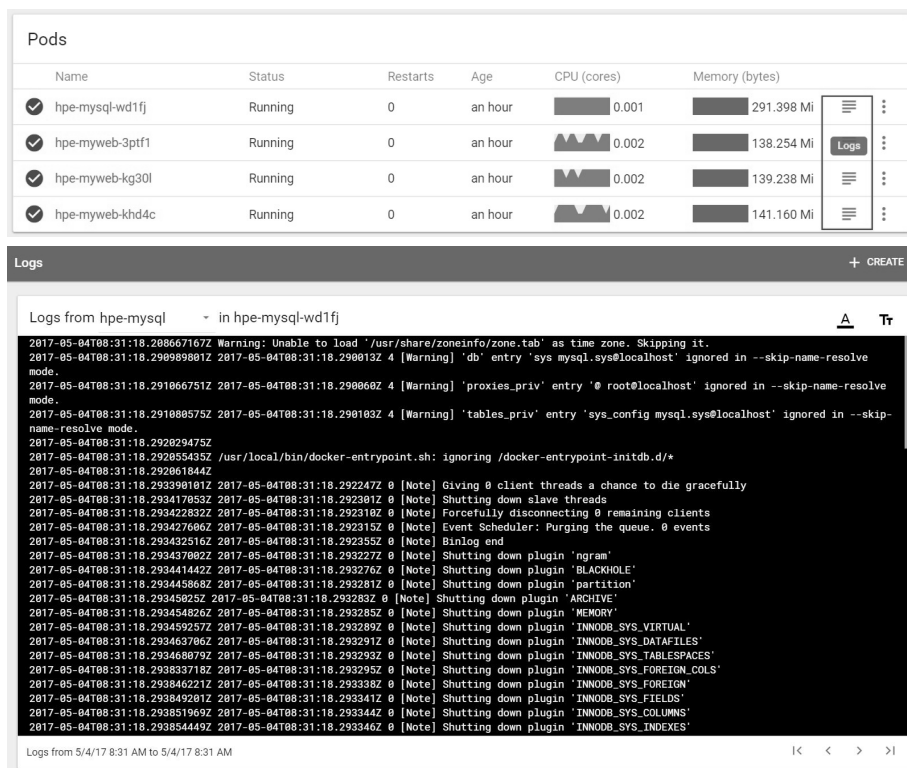
Services

Name	Labels	Cluster IP	Internal endpoints
✔ hpe-mysql	-	169.169.231.96	hpe-mysql:3306 TCP hpe-mysql:0 TCP
✔ hpe-myweb	-	169.169.103.90	hpe-myweb:8080 TCP hpe-myweb:30001 TCP

图 5.23 Service 列表

在 Pod 列表中，在各 Pod 右侧可以查看容器应用的日志，如图 5.24 所示。





**Pods**

Name	Status	Restarts	Age	CPU (cores)	Memory (bytes)	
hpe-mysql-wd1fj	Running	0	an hour	0.001	291.398 Mi	⋮
hpe-myweb-3ptf1	Running	0	an hour	0.002	138.254 Mi	⋮
hpe-myweb-kg30l	Running	0	an hour	0.002	139.238 Mi	⋮
hpe-myweb-khd4c	Running	0	an hour	0.002	141.160 Mi	⋮

**Logs** + CREATE

Logs from hpe-mysql - in hpe-mysql-wd1fj

```

2017-05-04T08:31:18.288667167Z Warning: Unable to load '/usr/share/zoneinfo/zone.tab' as time zone. Skipping it.
2017-05-04T08:31:18.290898801Z 2017-05-04T08:31:18.290813Z 4 [Warning] 'db' entry 'sys mysql.sys@localhost' ignored in --skip-name-resolve mode.
2017-05-04T08:31:18.291066751Z 2017-05-04T08:31:18.290868Z 4 [Warning] 'proxies_priv' entry ''@root@localhost' ignored in --skip-name-resolve mode.
2017-05-04T08:31:18.291088575Z 2017-05-04T08:31:18.290103Z 4 [Warning] 'tables_priv' entry 'sys_config mysql.sys@localhost' ignored in --skip-name-resolve mode.
2017-05-04T08:31:18.292029475Z
2017-05-04T08:31:18.292055435Z /usr/local/bin/docker-entrypoint.sh: ignoring /docker-entrypoint-initdb.d/*
2017-05-04T08:31:18.292061844Z
2017-05-04T08:31:18.293398101Z 2017-05-04T08:31:18.292247Z 0 [Note] Giving 0 client threads a chance to die gracefully
2017-05-04T08:31:18.293417053Z 2017-05-04T08:31:18.292301Z 0 [Note] Shutting down slave threads
2017-05-04T08:31:18.293422832Z 2017-05-04T08:31:18.292310Z 0 [Note] Forcefully disconnecting 0 remaining clients
2017-05-04T08:31:18.293427066Z 2017-05-04T08:31:18.292315Z 0 [Note] Event Scheduler: Purging the queue. 0 events
2017-05-04T08:31:18.293432515Z 2017-05-04T08:31:18.292355Z 0 [Note] Binlog end
2017-05-04T08:31:18.293437022Z 2017-05-04T08:31:18.292327Z 0 [Note] Shutting down plugin 'ngram'
2017-05-04T08:31:18.293441442Z 2017-05-04T08:31:18.292326Z 0 [Note] Shutting down plugin 'BLACKHOLE'
2017-05-04T08:31:18.293445868Z 2017-05-04T08:31:18.292321Z 0 [Note] Shutting down plugin 'partition'
2017-05-04T08:31:18.29345025Z 2017-05-04T08:31:18.292328Z 0 [Note] Shutting down plugin 'ARCHIVE'
2017-05-04T08:31:18.293454626Z 2017-05-04T08:31:18.292325Z 0 [Note] Shutting down plugin 'MEMORY'
2017-05-04T08:31:18.293459257Z 2017-05-04T08:31:18.292320Z 0 [Note] Shutting down plugin 'INNODB_SYS_VIRTUAL'
2017-05-04T08:31:18.293463706Z 2017-05-04T08:31:18.292321Z 0 [Note] Shutting down plugin 'INNODB_SYS_DATAFILES'
2017-05-04T08:31:18.293468097Z 2017-05-04T08:31:18.292322Z 0 [Note] Shutting down plugin 'INNODB_SYS_TABLESPACES'
2017-05-04T08:31:18.293472518Z 2017-05-04T08:31:18.292323Z 0 [Note] Shutting down plugin 'INNODB_SYS_FOREIGN_COLS'
2017-05-04T08:31:18.293476939Z 2017-05-04T08:31:18.292324Z 0 [Note] Shutting down plugin 'INNODB_SYS_FOREIGN'
2017-05-04T08:31:18.293481360Z 2017-05-04T08:31:18.292325Z 0 [Note] Shutting down plugin 'INNODB_SYS_FIELDS'
2017-05-04T08:31:18.293485781Z 2017-05-04T08:31:18.292326Z 0 [Note] Shutting down plugin 'INNODB_SYS_COLUMNS'
2017-05-04T08:31:18.293490202Z 2017-05-04T08:31:18.292327Z 0 [Note] Shutting down plugin 'INNODB_SYS_INDEXES'
  
```

Logs from 5/4/17 8:31 AM to 5/4/17 8:31 AM

图 5.24 查看容器应用的日志

### 5.1.12 Helm: Kubernetes 应用包管理工具

随着容器技术逐渐被企业接受，简单的应用在 Kubernetes 上已经能够便捷部署。但对于复杂的应用中间件，在 Kubernetes 上进行容器化部署并非易事，通常需要先研究 Docker 镜像的运行需求、环境变量等内容，并能为这些容器定制存储、网络等设置，最后设计和编写 Deployment、Configmap、Service 及 Ingress 等相关 yaml 配置文件，再提交给 Kubernetes 进行部署。这些复杂的过程将逐步被 Helm 应用包管理工具实现。

#### 1. Helm 概述

Helm 是一个由 CNCF 孵化和管理的项目，用于对需要在 Kubernetes 上部署的复杂应用进行定义、安装和更新。Helm 以 Chart 的方式对应用软件进行描述，可以方便地创建、版本化、共享和发布复杂的应用软件。

## 2. Helm 的主要概念

- ◎ **Chart**: 一个 Helm 包，其中包含了运行一个应用所需要的工具和资源定义，还可能包含 Kubernetes 集群中的服务定义，类似于 Homebrew 中的 formula、APT 中的 dpkg 或者 Yum 中的 RPM 文件。
- ◎ **Release**: 在 Kubernetes 集群上运行的一个 Chart 实例。在同一个集群上，一个 Chart 可以安装多次。例如有一个 MySQL Chart，如果想在服务器上运行两个 MySQL 数据库，就可以基于这个 Chart 安装两次。每次安装都会生成新的 Release，会有独立的 Release 名称。
- ◎ **Repository**: 用于存放和共享 Chart 的仓库。

简单来说，Helm 整个系统的主要任务就是，在仓库中查找需要的 Chart，然后将 Chart 以 Release 的形式安装到 Kubernetes 集群中。

## 3. Helm 的安装

Helm 由以下两个组件组成。

- ◎ **HelmClient**: 客户端，拥有对 Repository、Chart、Release 等对象的管理能力。
- ◎ **TillerServer**: 负责客户端指令和 Kubernetes 集群之间的交互，根据 Chart 定义，生成和管理各种 Kubernetes 的资源对象。

1) HelmClient 客户端：可以通过二进制文件或脚本方式进行安装。

通过二进制文件方式进行安装：从 <https://github.com/kubernetes/helm/releases> 下载二进制文件，解压并复制到执行目录即可。

通过脚本方式进行安装：

```
# curl https://raw.githubusercontent.com/kubernetes/helm/master/scripts/get | bash
```

2) TillerServer 的安装

(1) 对 TillerServer 的安装可以使用 `helm init` 命令进行（官方推荐），这一命令会在 `kubectl` 当前 context 指定集群内的 `kube-system` 命名空间创建一个 `Deployment` 和一个 `Service`，运行 TillerServer 服务。

Deployment 中使用的镜像是 `gcr.io/kubernetes-helm/tiller:v[helm-version]`，Helm 版本可以使用 `helm version` 命令获得。如果环境无法连接互联网获取该镜像，则可以先通过一台能够联网的服务器下载这个镜像并保存到镜像私库，利用 `helm init` 子命令的 `--tiller-image` 参数来指定私库中的镜像来执行初始化过程。

安装结束之后，用 `helm version` 命令来验证安装情况，一切正常的话，会分别显示 Tiller 和 Helm 的版本信息。一个常见的问题是，Tiller 部分显示一个错误信息：“`uid : unable to do port forwarding: socat not found.`”，这是因为所在节点没有 `socat`，无法进行端口转发造成的，在主机上安装 `socat` 软件即可。

(2) 对 TillerServer 的安装还可以在本地进行，在服务器本地直接运行 Tiller，这种安装方式需要让 Helm 指定要连接的服务地址，有以下两种方法。

- ◎ 使用 `--host` 指定 Tiller 运行的监听地址。
- ◎ 设置 `HELM_HOST` 环境变量。

需要注意的是，Tiller 仍会使用 `kubectl` 配置中的 `context` 连接 Kubernetes 集群。

## 4. Helm 的使用

下面介绍 Helm 的常见用法，包括搜索 Chart、安装 Chart、自定义 Chart 配置、更新或回滚 Release、删除 Release、创建自定义 Chart、搭建私有仓库等。

### 1) helm search: 搜索可用的 Chart

Helm 初始化完成之后，默认配置为使用官方的 Kubernetes Chart 仓库。官方仓库包含大量的经过组织和持续维护的 Chart，这个仓库通常命名为 `stable`。

使用 `helm search` 命令来查找可安装的 Chart：

```
$ helm search
NAME          VERSION   DESCRIPTION
local/gitlab   0.1.3     A Helm chart for Kubernetes
local/grafana  0.1.0
local/influxdb 0.1.3     Fast, reliable, scalable, and easy to use open-...
local/jenkins  0.1.8     A Helm chart for Kubernetes
.....
```

在没有进行过滤的情况下，`helm search` 会显示所有可用 Chart，可以使用参数进行过滤：

```
$ helm search mysql
NAME          VERSION   DESCRIPTION
local/mysql   0.2.6     Fast, reliable, scalable, and easy to use open-...
stable/mysql  0.2.6     Fast, reliable, scalable, and easy to use open-...
stable/percona 0.1.0     free, fully compatible, enhanced, open source d...
stable/mariadb 0.6.0     Fast, reliable, scalable, and easy to use open-...
```

为什么列表中会有 MariaDB？因为 MariaDB 的描述信息中包含了 `mysql` 关键字。可以使用 `helm inspect <chart_name>` 命令查看 Chart 的详细信息：

```
$ helm inspect stable/mariadb
description: Fast, reliable, scalable, and easy to use open-source relational
```

```
database
  system. MariaDB Server is intended for mission-critical, heavy-load production
systems
  as well as for embedding into mass-deployed software.
engine: gotpl
home: https://mariadb.org
icon: https://bitnami.com/assets/stacks/mariadb/img/mariadb-stack-220x234.png
keywords:
- mariadb
- mysql
- database
- sql
...
```

找到需要安装的 Chart 之后，就可以进行安装了。

## 2) helm install: 安装 Chart

使用 `helm install` 命令进行应用的安装，最简单的参数就是 Chart 的名称。下面以 MariaDB 为例：

```
$ helm install stable/mariadb
NAME: womping-bobcat
LAST DEPLOYED: Wed May 31 21:13:58 2017
NAMESPACE: default
STATUS: DEPLOYED

RESOURCES:
==> v1/Service
NAME                                CLUSTER-IP    EXTERNAL-IP  PORT(S)    AGE
womping-bobcat-mariadb 10.31.255.15  <none>       3306/TCP   1s

==> v1beta1/Deployment
NAME                                DESIRED  CURRENT  UP-TO-DATE  AVAILABLE  AGE
womping-bobcat-mariadb 1         1         1             0           1s

==> v1/Secret
NAME                                TYPE      DATA  AGE
womping-bobcat-mariadb Opaque    2       1s

==> v1/ConfigMap
NAME                                DATA  AGE
womping-bobcat-mariadb 1       1s

==> v1/PersistentVolumeClaim
NAME                                STATUS  VOLUME    CAPACITY  ACCESSMODES  STORAGECLASS
...
```

至此，MariaDB 就安装完成了。可以看到系统创建了一个新的名为 `womping-bobcat` 的 Release 对象，这个名称可以在 `helm install` 命令中使用 `--name` 参数进行修改。

在安装过程中，Helm 客户端会输出一些有用的信息，例如 Release 的状态，以及额外的配置步骤等。

Helm 不会等待所有创建过程的完成。这是因为有些 Chart 的 Docker 镜像较大，会消耗很长的时间进行下载和创建。

在 `helm install` 过程中，可以使用 `helm status` 命令来跟踪 release 的状态：

```
$ helm status womping-bobcat [21:28:26]
LAST DEPLOYED: Wed May 31 21:13:58 2017
NAMESPACE: default
STATUS: DEPLOYED

RESOURCES:
==> v1/Secret
NAME                                TYPE      DATA  AGE
womping-bobcat-mariadb              Opaque    2       14m

==> v1/ConfigMap
NAME                                DATA  AGE
womping-bobcat-mariadb              1       14m
.....
```

在成功安装 Chart 后，系统会在 `kube-system` 命名空间内创建一个 configmap 用于保存 Release 对象的数据：

```
$ kubectl get configmap
NAMESPACE   NAME                                DATA  AGE
default     womping-bobcat-mariadb              1       18m
```

### 3. 自定义 Chart 的配置

前面介绍的安装过程使用的是 Chart 的默认配置。然而在很多情况下，我们会希望使用自定义的配置进行应用的部署。

首先，用 `helm inspect` 命令查看 Chart 的可配置内容：

```
$ helm inspect stable/mariadb
.....
## Bitnami MariaDB image version
## ref: https://hub.docker.com/r/bitnami/mariadb/tags/
##
## Default: none
image: bitnami/mariadb:10.1.22-r1

## Specify an imagePullPolicy (Required)
```

```
## It's recommended to change this to 'Always' if the image tag is 'latest'
## ref: http://kubernetes.io/docs/user-guide/images/#updating-images
imagePullPolicy: IfNotPresent

## Specify password for root user
## ref:
https://github.com/bitnami/bitnami-docker-mariadb/blob/master/README.md#setting-
the-root-password-on-first-run
##
# mariadbRootPassword:

## Create a database user
## ref:
https://github.com/bitnami/bitnami-docker-mariadb/blob/master/README.md#creating
-a-database-user-on-first-run
##
# mariadbUser:
# mariadbPassword:
.....
```

用户可以编写一个 `yaml` 配置文件来覆盖上面这些设置，然后利用这一文件来给安装过程提供配置。例如，我们可以自定义额外的两个配置文件 `config.yaml` 和 `config2.yaml`，用于 `helm install` 安装 MariaDB 之后，在 MariaDB 启动时自动创建名为 `firstdb` 和 `seconddb` 的数据库，并且设置 `root` 用户的密码：

```
$ echo 'mariadbDatabase: firstdb' > config.yaml
$ echo 'mariadbRootPassword: abcdefgh' > config2.yaml
$ echo 'mariadbDatabase: seconddb' >> config2.yaml
$ helm install stable/mariadb -f config.yaml -f config2.yaml
```

安装完成之后，可以登录到 MariaDB Pod 中查看数据库是否已经创建成功。

自定义 Chart 的配置有两种方法。

- ◎ `--values` 或者 `-f`：使用 `yaml` 配置文件进行参数配置，可以设置多个文件，最后一个优先生效。多个文件中重复的 `value` 会进行覆盖操作，不同的 `value` 会叠加生效。上面的例子使用的就是这种方式。
- ◎ `--set`：在命令行直接设置参数。

如果同时使用两个参数，则 `--set` 会以高优先级合并到 `--values` 中。

## 5. `--set` 的格式和限制

`--set` 参数可以使用零或多个名称/值的组合。最简单的方式是 `--set name=value`，`yaml` 中的等效描述是：

```
name: value
```

多个值可以使用逗号进行分隔，例如`--set a=b,c=d`的 yml 等效为下面的描述：

```
a: b
c: d
```

还可以用来表达多层结构的变量`--set outer.inner=value`：

```
outer:
  inner: value
```

大括号（{}）可以用来表达列表数据，例如`--set name={a,b,c}`会翻译成：

```
name:
- a
- b
- c
```

有时需要在`--set`时使用一些特殊字符，这里可以使用斜线进行转义，比如`--set name=value1\,value2`。类似地，可以对点符号“.”进行转义，这样 Chart 使用 `toYaml` 函数解析注解、标签或者 node selector 时就方便了，例如：`--set nodeSelector. "kubernetes\io/ role"=master`。

尽管如此，`--set`语法的表达能力依然无法和 yml 相提并论，尤其是在处理集合时。目前没有方法能够完成注入“把列表中第 3 个元素设置为 XXX”这样的描述。

## 6. 更多的安装方法

`helm install` 能通过多种安装源进行安装。

- ◎ 上面用到的 Chart 仓库。
- ◎ 本地的 Chart 压缩包（`helm install foo-0.1.1.tgz`）。
- ◎ 一个 Chart 目录（`helm install path/to/foo`）。
- ◎ 一个完整的 URL（`helm install https://example.com/charts/foo-1.2.3.tgz`）。

## 7. helm upgrade 和 helm rollback：应用的更新或回滚

当一个 Chart 发布新版本或者需要修改一个 Release 的配置时，就需要使用 `helm upgrade` 命令了。

`helm upgrade` 会利用用户提供的更新信息来对 Release 进行更新。因为 Kubernetes Chart 可能会有很大的规模或者相对复杂的关系，Helm 会尝试进行最小影响范围的更新，只更新相对于上一个 Release 来说发生变化的内容。

例如我们要更新一个 Release 的资源限制，创建 `config3.yaml` 配置文件，内容如下：

```
resources:
  requests:
    memory: 256Mi
    cpu: 500m
```

使用 **upgrade** 命令完成更新：

```
$ helm upgrade -f config3.yaml nomadic-terrier stable/mariadb
```

看到更新提示之后，我们可以用 **Helm** 的 **list** 指令查看 **Release** 的信息，会发现 **revision** 一列发生了变化。接下来使用 **kubect**l 的 **get pods**，**deploy** 指令，可以看到 **pod** 已经被更新；如果使用 **kubect**l **describe deploy** 指令，则还会看到 **deployment** 的更新过程和一系列的 **ScalingReplicaSet** 事件。

如果对更新后的 **Release** 不满意，则可以使用 **helm rollback** 命令对 **Release** 进行回滚，例如：

```
$ helm rollback nomadic-terrier 2
```

这个命令将把名为 **nomadic-terrier** 的 **Release** 回滚到版本 2。

命令执行之后，同样可以使用前面提到的几个查询指令，会看到类似的结果。

最后，我们可以使用 **helm history <release\_name>** 命令来查看一个 **Release** 的变更历史。

## 8. helm install/upgrade/rollback 命令的常用参数

**Helm** 有很多参数可以帮助用户来指导命令的行为。本节介绍一些常用参数，用户可以使用 **helm <command> help** 命令来获取所有参数的列表。

- ◎ **--timeout**：等待 **Kubernetes** 命令完成的时间，单位是 s，默认值为 300，也就是 5min。
- ◎ **--wait**：等待 **Pod**，直到其状态变为 **ready**，**PVC** 完成绑定，**Deployment** 完成其最低就绪要求的 **Pod** 创建，并且服务有了 **IP** 地址，才认为 **Release** 创建成功。这一等待过程会一直持续到超过 **--timeout**，超时后这一 **Release** 被标记为 **FAILED**（注意：当 **Deployment** 的 **replicas** 被设置为 1，同时滚动更新策略的 **maxUnavailable** 不为 0 时，**--wait** 会因为最小就绪 **Pod** 数量达成而返回 **ready** 状态）。
- ◎ **--no-hooks**：该命令会跳过 **Hook** 执行。
- ◎ **--recreate-pods**：会引起所有 **pod** 的重建（**Deployment** 所属的 **Pod** 除外）。

## 9. helm delete：删除一个 Release

执行 **helm delete** 命令可以删除一个 **Release**，例如使用 **helm delete happy-panda** 会从集群中删除名为“**happy-panda**”的 **Release**。

可以使用 **helm list** 命令列出集群中部署的 **Release**。如果给 **list** 加上 **--deleted** 参数，则会列



出所有删除的 Release; --all 参数会列出所有的 Release, 包含删除的、现存的及失败的 Release。

正因为 Helm 会保存所有被删除 Release 的信息, 所以 Release 的名字是不可复用的 (如果坚持复用, 则可以使用 --replace 参数, 这一操作不建议在生产环境中使用), 这样被删除的 Release 也可以被回滚, 甚至重新激活。

## 10. helm repo: 仓库的使用

我们在前面使用的 Chart 来自于 stable 仓库。我们也可以对 helm 进行配置, 让其使用其他仓库。Helm 在 helm repo 命令中提供了很多仓库相关的工具。

- ◎ helm repo list: 列出所有仓库。
- ◎ helm repo add: 添加仓库, 例如从 repo\_url 添加名为 dev 的仓库 helm repo add dev http://<repo\_url>/dev-charts。
- ◎ helm repo update: 更新仓库中的 Chart 信息。

## 11. 自定义 Chart

用户可以将自己的应用定义为 Chart 并进行打包部署, 本节对其进行简单介绍, 详细的开发指南参见 <https://docs.helm.sh/developing-charts>。

自定义 Chart 需要使用符合 Helm 规范的一组目录和配置文件来完成。

## 12. Chart 目录结构和配置文件的说明

Chart 是一个包含一系列文件的目录。目录的名字就是 Chart 的名字 (不包含版本信息), 例如一个 WordPress 的 Chart 就会存储在 wordpress 目录中。

目录中的文件结构如下:

```
wordpress/
  Chart.yaml      # 用于描述 Chart 信息的 YAML 文件
  LICENSE         # 可选: Chart 的许可信息
  README.md       # 可选: README 文件
  values.yaml     # 默认的配置值
  charts/         # 可选: 包含该 Chart 所依赖的其他 Chart, 依赖管理推荐采用
requirements.yaml 文件来进行
  templates/      # 可选: 结合 values.yaml, 能够生成 Kubernetes 的 manifest 文件
  templates/NOTES.txt # 可选: 文本文件, 用法描述
```

charts/子目录和 requirements.yaml 的区别在于, 前者需要提供整个 Chart 的文件, 后者仅需要注明依赖 Chart 的仓库信息, 例如一个 requirements.yaml 可以定义为:

```
dependencies:
- name: apache
  version: 1.2.3
  repository: http://example.com/charts
- name: mysql
  version: 3.2.1
  repository: http://another.example.com/charts
```

### 13. Chart.yaml 文件说明

Chart.yaml 文件（注意首字母大写）是个必要文件，包含如下内容。

- ◎ **name:** Chart 名，必选。
- ◎ **version:** SemVer 2 规范的版本号，必选。
- ◎ **description:** 项目的描述，可选。
- ◎ **keywords:** 一个用于描述项目的关键字列表，可选。
- ◎ **home:** 项目的主页，可选。
- ◎ **sources:** 一个 URL 列表，说明项目的源代码位置，可选。
- ◎ **maintainers:** 维护者列表，可选。
- ◎ **name:** 管理员名字，必选。
- ◎ **email:** 管理员邮件，必选。
- ◎ **engine:** 模板引擎名称，默认是 gotpl，可选。
- ◎ **icon:** 一个指向 svg 或 png 图像的 URL，作为 Chart 的图标，可选。
- ◎ **appVersion:** Chart 中包含的应用的版本，无须遵循 SemVer 规范，可选。
- ◎ **deprecated:** 布尔值，该 Chart 是否标注为“弃用”，可选。
- ◎ **tillerVersion:** 可选，该 Chart 所需的 Tiller 版本。取值应该是一个 SemVer 的范围，例如“>2.0.0”。

### 14. 快速制作自定义的 Chart

同其他软件开发过程一样，快速制作一个简单 Chart 的方法，就是从其他项目中复制并修改。例如我们要简单地改写前面 MariaDB 的 Chart，令其使用本地的私有镜像仓库，可以按照如下步骤进行。

- ◎ **下载 Chart:** 使用 `helm fetch stable/mariadb` 命令下载这一 Chart 的压缩包。

- ◎ 编辑 Chart。
- ◎ 利用 tar 解压之后，我们将目录重新命名为 mymariadb。
- ◎ 修改 templates 中的 deployment.yaml，简单地将其中的 image 字段硬编码为需要的镜像（当然不推荐这种用法，可以继续以变量的方式在 values 中进行设置）。
- ◎ 将 Chart.yaml 中的版本号修改为 0.1.1，name 为 mymariadb。
- ◎ 使用 helm package mymariadb 打包 Chart，会生成一个名为 mymariadb-0.1.1.tgz 的压缩包。
- ◎ 安装 Chart：通过 helm install mymariadb-0.1.1.tgz 命令即可将我们“新”生成的 Chart 安装到集群中。

## 15. 搭建私有 Repository

自建 Chart 之后，自然需要搭建私有仓库。下面使用 Apache 来搭建一个简单的 Chart 私有仓库，并将刚才新建的 mymariadb Chart 保存到私有仓库中。详情可参考 <https://docs.helm.sh/developing-charts/#chart-repo-guide>。

Chart 仓库主要由前面提到的 Chart 压缩包和索引文件构成，通过 HTTP/HTTPS 对外提供服务。这里我们使用一个 Apache 应用来提供仓库服务。Apache 的设置如下。

- ◎ Apache 使用 /var/web/repo 目录进行仓库的存储。
- ◎ 使用 http://127.0.0.1/repo 网址提供访问服务。

将前面生成的 mymariadb-0.1.1.tgz 文件复制到仓库的 /var/web/repo 目录中。

接下来使用 helm repo index /var/web/repo --url http://127.0.0.1/repo 命令，Helm 将自动根据目录中的内容创建索引。命令执行完毕后，可以看到目录下多出了一个 index.yaml 文件。

最后启动 Web Server。

为了能够使用这个私有仓库，需要将这个新的仓库地址加入到 Helm 配置中：

```
$ helm repo add localhost http://127.0.0.1/repo
```

再次运行 helm search mysql 命令，就会看到 Chart 列表中多出来的 localhost/mymariadb 项目，也就是我们的新仓库中的 Chart。

现在就可以使用 helm install localhost/mymariadb 来安装私有仓库中的 Chart 了。

## 5.2 Trouble Shooting 指导

本节将对 Kubernetes 集群中常见问题的排查方法进行说明。

为了跟踪和发现 Kubernetes 集群中运行的容器应用出现的问题，常用的查错方法如下。

首先，查看 Kubernetes 对象的当前运行时信息，特别是与对象关联的 Event 事件。这些事件记录了相关主题、发生时间、最近发生时间、发生次数及事件原因等，对排查故障非常有价值。此外，通过查看对象的运行时数据，我们还可以发现参数错误、关联错误、状态异常等明显问题。由于 Kubernetes 中多种对象相互关联，因此，这一步可能会涉及多个相关对象的排查问题。

其次，对于服务、容器的问题，则可能需要深入容器内部进行故障诊断，此时可以通过查看容器的运行日志来定位具体问题。

最后，对于某些复杂问题，比如 Pod 调度这种全局性的问题，可能需要结合集群中每个节点上的 Kubernetes 服务日志来排查。比如搜集 Master 上 kube-apiserver、kube-schedule、kube-controller-manager 服务的日志，以及各个 Node 节点上的 kubelet、kube-proxy 服务的日志，综合判断各种信息，我们就能找到问题的成因并解决问题。

### 5.2.1 查看系统 Event 事件

在 Kubernetes 集群中创建了 Pod 之后，我们可以通过 `kubectl get pods` 命令查看 Pod 列表，但该命令能够显示的信息很有限。Kubernetes 提供了 `kubectl describe pod` 命令来查看一个 Pod 的详细信息。

```
$ kubectl describe pod redis-master-bobr0
Name:                                Redis-master-bobr0
Namespace:                           default
Image(s):                            kubeguide/Redis-master
Node:                                k8s-node-1/192.168.18.3
Labels:                               name=Redis-master,role=master
Status:                               Running
Reason:
Message:
IP:                                  172.17.0.58
Replication Controllers:              Redis-master (1/1 replicas created)
Containers:
  master:
    Image:                            kubeguide/Redis-master
```

```

Limits:
  cpu:          250m
  memory:       64Mi
State:         Running
  Started:      Fri, 21 Aug 2015 14:45:37 +0800
Ready:         True
Restart Count: 0
Conditions:
  Type          Status
  Ready         True

Events:

```

Reason	FirstSeen	Message	LastSeen	Count	From	SubobjectPath
	Fri, 21 Aug 2015 14:45:36 +0800	{kubelet k8s-node-1} implicitly required container image "myregistry:5000/google_containers/pause:latest" already present on machine	Fri, 21 Aug 2015 14:45:36 +0800	1	pulled	Pod
	Fri, 21 Aug 2015 14:45:37 +0800	{kubelet k8s-node-1} implicitly required container image "myregistry:5000/google_containers/pause:latest" already present on machine	Fri, 21 Aug 2015 14:45:37 +0800	1	created	Created
	Fri, 21 Aug 2015 14:45:37 +0800	{kubelet k8s-node-1} implicitly required container image "myregistry:5000/google_containers/pause:latest" already present on machine	Fri, 21 Aug 2015 14:45:37 +0800	1	started	Started
	Fri, 21 Aug 2015 14:45:37 +0800	{kubelet k8s-node-1} spec.containers{master} Created with docker id 1e746245f768	Fri, 21 Aug 2015 14:45:37 +0800	1	created	
	Fri, 21 Aug 2015 14:45:37 +0800	{kubelet k8s-node-1} spec.containers{master} Started with docker id 1e746245f768	Fri, 21 Aug 2015 14:45:37 +0800	1	started	
	Fri, 21 Aug 2015 14:45:37 +0800	{scheduler } Redis-master-bobr0 to k8s-node-1	Fri, 21 Aug 2015 14:45:37 +0800	1	scheduled	Successfully assigned

该命令除了显示 Pod 创建时的配置定义、状态等信息，还显示了与该 Pod 相关的最近的 Event 事件，事件信息对于查错非常有用。如果某个 Pod 一直处于 Pending 状态，则我们通过 `kubectl describe` 命令就能了解到失败的具体原因。例如，从 Event 事件中我们可能获知 Pod 失败的原因有以下几种。

- ◎ 没有可用的 Node 以供调度。
- ◎ 开启了资源配额管理并且当前 Pod 的目标节点上恰好没有可用的资源。
- ◎ 正在下载镜像。

`kubectl describe` 命令还可用于查看其他 Kubernetes 对象，包括 Node、RC、Service、Namespace、Secrets 等，对于每一种对象都会显示相关的其他信息。

例如，查看一个服务的详细信息：

```
$ kubectl describe service redis-master
Name:                Redis-master
Namespace:           default
Labels:              name=Redis-master
Selector:            name=Redis-master
Type:                ClusterIP
IP:                  169.169.208.57
Port:                <unnamed>      6379/TCP
Endpoints:           172.17.0.58:6379
Session Affinity:    None
No events.
```

如果查看的对象属于某个特定的 namespace，则需要加上 `--namespace=<namespace>` 进行查询。例如：

```
$ kubectl get service kube-dns --namespace=kube-system
```

## 5.2.2 查看容器日志

在需要排查容器内部应用程序生成的日志时，我们可以使用 `kubectl logs <pod_name>` 命令：

```
$ kubectl logs redis-master-bobr0
[1] 21 Aug 06:45:37.781 * Redis 2.8.19 (00000000/0) 64 bit, stand alone mode,
port 6379, pid 1 ready to start.
[1] 21 Aug 06:45:37.781 # Server started, Redis version 2.8.19
[1] 21 Aug 06:45:37.781 # WARNING overcommit_memory is set to 0! Background save
may fail under low memory condition. To fix this issue add 'vm.overcommit_memory =
1' to /etc/sysctl.conf and then reboot or run the command 'sysctl
vm.overcommit_memory=1' for this to take effect.
[1] 21 Aug 06:45:37.782 # WARNING you have Transparent Huge Pages (THP) support
enabled in your kernel. This will create latency and memory usage issues with Redis.
To fix this issue run the command 'echo never > /sys/kernel/mm/transparent_hugepage/
enabled' as root, and add it to your /etc/ rc.local in order to retain the setting
after a reboot. Redis must be restarted after THP is disabled.
[1] 21 Aug 06:45:37.782 # WARNING: The TCP backlog setting of 511 cannot be enforced
because /proc/sys/net/core/somaxconn is set to the lower value of 128.
```

如果在一个 Pod 中包含多个容器，则需要通过 `-c` 参数指定容器的名称来进行查看，例如：

```
kubectl logs <pod_name> -c <container_name>
```

这个命令与在 Pod 的宿主机上运行 `docker logs <container_id>` 的效果是一样的。

容器中应用程序生成的日志与容器的生命周期是一致的，所以在容器被销毁之后，容器内部的文件也会被丢弃，包括日志等。如果需要保留容器内应用程序生成的日志，则一方面可以

使用挂载的 Volume（存储卷）将容器产生的日志保存到宿主机，另一方面可以通过一些工具对日志进行采集，包括 Fluentd、Elasticsearch 等开源软件。

### 5.2.3 查看 Kubernetes 服务日志

如果在 Linux 系统上进行安装，并且使用 systemd 系统来管理 Kubernetes 服务，那么 systemd 的 journal 系统会接管服务程序的输出日志。在这种环境中，可以通过使用 `systemd status` 或 `journalctl` 工具来查看系统服务的日志。

例如，使用 `systemctl status` 命令查看 `kube-controller-manager` 服务的日志：

```
# systemctl status kube-controller-manager -l
kube-controller-manager.service - Kubernetes Controller Manager
   Loaded: loaded (/usr/lib/systemd/system/kube-controller-manager.service;
   enabled)
   Active: active (running) since Fri 2015-08-21 18:36:29 CST; 5min ago
     Docs: https://github.com/GoogleCloudPlatform/kubernetes
    Main PID: 20339 (kube-controller)
      CGroup: /system.slice/kube-controller-manager.service
              └─20339 /usr/bin/kube-controller-manager --logtostderr=false --v=4
--master=http://kubernetes-master:8080 --log_dir=/var/log/kubernetes
```

```
Aug 21 18:36:29 kubernetes-master systemd[1]: Starting Kubernetes Controller
Manager...
Aug 21 18:36:29 kubernetes-master systemd[1]: Started Kubernetes Controller Manager.
```

使用 `journalctl` 命令查看：

```
# journalctl -u kube-controller-manager
-- Logs begin at Mon 2015-08-17 16:43:22 CST, end at Fri 2015-08-21 18:36:29 CST.
--
Aug 17 16:44:14 kubernetes-master systemd[1]: Starting Kubernetes Controller
Manager...
Aug 17 16:44:14 kubernetes-master systemd[1]: Started Kubernetes Controller Manager.
```

如果不使用 systemd 系统接管 Kubernetes 服务的标准输出，则也可以通过日志相关的启动参数来指定日志的存放目录。

- ◎ `--logtostderr=false`：不输出到 `stderr`。
- ◎ `--log-dir=/var/log/kubernetes`：日志的存放目录。
- ◎ `--alsologtostderr=false`：设置为 `true` 则表示将日志输出到文件时也输出到 `stderr`。
- ◎ `--v=0`：glog 日志级别。
- ◎ `--vmodule=gfs*=2,test*=4`：glog 基于模块的详细日志级别。

在 `--log_dir` 设置的目录中可以查看各服务进程生成的日志文件，日志文件的数量和大小依赖于日志级别的设置。例如 `kube-controller-manager` 可能生成的几个日志文件如下。

- ◎ `kube-controller-manager.ERROR`。
- ◎ `kube-controller-manager.INFO`。
- ◎ `kube-controller-manager.WARNING`。
- ◎ `kube-controller-manager.kubernetes-master.unknownuser.log.ERROR.20150930-173939.9847`。
- ◎ `kube-controller-manager.kubernetes-master.unknownuser.log.INFO.20150930-173939.9847`。
- ◎ `kube-controller-manager.kubernetes-master.unknownuser.log.WARNING.20150930-173939.9847`。

在大多数情况下，我们从 `WARNING` 和 `ERROR` 级别的日志中就能找到问题的原因，但有时还是需要排查 `INFO` 级别的日志甚至 `DEBUG` 级别的详细日志。此外，`etcd` 服务也属于 Kubernetes 集群中的重要组成部分，所以它的日志也不能忽略。

如果某个 Kubernetes 对象存在问题，则我们可以用这个对象的名字作为关键字搜索 Kubernetes 的日志来发现和解决问题。在大多数情况下，我们平常所遇到的主要是与 Pod 对象相关的问题，比如无法创建 Pod、Pod 启动后就停止或者 Pod 副本无法增加等。此时，我们可以先确定 Pod 在哪个节点上，然后登录这个节点，从 `kubelet` 的日志中查询该 Pod 的完整日志，然后进行问题排查。对于与 Pod 扩容相关或者与 RC 相关的问题，则很可能在 `kube-controller-manager` 及 `kube-scheduler` 的日志中找出问题的关键点。

另外，`kube-proxy` 经常被我们忽视，因为即使它意外地被停止，Pod 的状态也是正常的，但会导致某些服务访问异常。这些错误通常与每个节点上的 `kube-proxy` 服务有着密切的关系。遇到这些问题时，首先要排查 `kube-proxy` 服务的日志，同时排查防火墙服务，特别是要留意防火墙中是否有人为添加的可疑规则。

## 5.2.4 常见问题

本节对 Kubernetes 系统中的一些常见问题及解决方法进行说明。

### 1. 由于无法下载 pause 镜像导致 Pod 一直处于 Pending 的状态

以 `redis-master` 为例，使用如下配置文件 `redis-master-controller.yaml` 创建 RC 和 Pod：

```
apiVersion: v1
kind: ReplicationController
metadata:
```



```

name: redis-master
labels:
  name: redis-master
spec:
  replicas: 1
  selector:
    name: redis-master
  template:
    metadata:
      labels:
        name: redis-master
    spec:
      containers:
        - name: master
          image: kubeguide/redis-master
          ports:
            - containerPort: 6379

```

执行 `kubectl create -f redis-master-controller.yaml` 成功。但在查看 Pod 时，发现其总是无法处于 **Running** 状态。通过 `kubectl get pods` 命令可以看到：

```

$ kubectl get pods
NAME                READY   STATUS              RESTARTS   AGE
redis-master-6yy7o  0/1     Image: kubeguide/redis-master is ready, container
is creating         0       5m

```

进一步使用 `kubectl describe pod redis-master-6yy7o` 命令查看该 Pod 的详细信息：

```

$ kubectl describe pod redis-master-6yy7o
Name:                redis-master-6yy7o
Namespace:           default
Image(s):             kubeguide/redis-master
Node:                127.0.0.1/127.0.0.1
Labels:              name=redis-master
Status:            Pending
Reason:
Message:
IP:
Replication Controllers:  redis-master (1/1 replicas created)
Containers:
  master:
    Image:            kubeguide/redis-master
    State:          Waiting
    Reason:           Image: kubeguide/redis-master is ready, container is
creating
    Ready:            False
    Restart Count:    0
Conditions:

```

```

Type           Status
Ready          False
Events:
  FirstSeen    LastSeen    Count  From              SubobjectPath
Reason    Message
  Thu, 24 Sep 2015 19:19:25 +0800    Thu, 24 Sep 2015 19:25:58 +0800  3
{kubelet 127.0.0.1}    failedSync Error syncing pod, skipping: image pull failed
for gcr.io/google_containers/pause-amd64:3.0, this may be because there are no
credentials on this request. details: (API error (500): invalid registry endpoint
https://gcr.io/v0/: unable to ping registry endpoint https://gcr.io/v0/v2 ping
attempt failed with error: Get https://gcr.io/v2/: dial tcp 173.194.196.82:443:
connection refused v1 ping attempt failed with error: Get https://gcr.io/v1/_ping:
dial tcp 173.194.79.82:443: connection refused. If this private registry supports
only HTTP or HTTPS with an unknown CA certificate, please add `--insecure-registry
gcr.io` to the daemon's arguments. In the case of HTTPS, if you have access to the
registry's CA certificate, no need for the flag; simply place the CA certificate at
/etc/docker/certs.d/gcr.io/ca.crt)
  Thu, 24 Sep 2015 19:19:25 +0800    Thu, 24 Sep 2015 19:25:58 +0800  3
{kubelet 127.0.0.1}    implicitly required container POD failed Failed to pull
image "gcr.io/google_containers/pause-amd64:3.0": image pull failed for
gcr.io/google_containers/pause:0.8.0, this may be because there are no credentials
on this request. details: (API error (500): invalid registry endpoint https://
gcr.io/v0/: unable to ping registry endpoint https://gcr.io/v0/v2 ping attempt failed
with error: Get https://gcr.io/v2/: dial tcp 173.194.196.82:443: connection refused
v1 ping attempt failed with error: Get https://gcr.io/v1/_ping: dial tcp 173.194.79.82:
443: connection refused. If this private registry supports only HTTP or HTTPS with
an unknown CA certificate, please add `--insecure-registry gcr.io` to the daemon's
arguments. In the case of HTTPS, if you have access to the registry's CA certificate,
no need for the flag; simply place the CA certificate at
/etc/docker/certs.d/gcr.io/ca.crt
```

可以看到，该 Pod 的状态为 Pending，从 Message 部分显示的信息可以看出其原因是 image pull failed for gcr.io/google\_containers/pause-amd64:3.0, 说明系统在创建 Pod 时无法从 gcr.io 下载 pause 镜像，所以导致创建 Pod 失败。

解决方法如下。

(1) 如果服务器可以访问 Internet，并且不希望使用 HTTPS 的安全机制来访问 gcr.io，则可以在 Docker Daemon 的启动参数中加上--insecure-registry gcr.io 来表示可以进行匿名下载。

(2) 如果 Kubernetes 的集群环境在内网环境中，无法访问 gcr.io 网站，则可以先通过一台能够访问 gcr.io 的机器将 pause 镜像下载下来，导出后，再导入内网的 Docker 私有镜像库中，并在 kubelet 的启动参数中加上--pod\_infra\_container\_image，配置为：

```
--pod_infra_container_image=<docker_registry_ip>:<port>/google_containers/pa
use-amd64:3.0
```

之后重新创建 `redis-master` 即可正确启动 Pod 了。

注意，除了 `pause` 镜像，其他 Docker 镜像也可能存在无法下载的情况，与上述情况类似，很可能也是网络配置使得镜像无法下载，解决方法同上。

## 2. Pod 创建成功，但状态始终不是 Ready，且 RESTARTS 的数量持续增加

在创建了一个 RC 之后，通过 `kubecttl get pods` 命令查看 Pod，发现如下情况：

```
.....
$ kubecttl get pods
NAME          READY    STATUS    RESTARTS   AGE
zk-bg-ri3ru   0/1      Running   3           37s
.....
$ kubecttl get pods
NAME          READY    STATUS    RESTARTS   AGE
zk-bg-ri3ru   0/1      Running   5           1m
.....
$ kubecttl get pods
NAME          READY    STATUS    RESTARTS   AGE
zk-bg-ri3ru   0/1      ExitCode:0 6           1m
.....
$ kubecttl get pods
NAME          READY    STATUS    RESTARTS   AGE
zk-bg-ri3ru   0/1      Running   7           1m
```

可以看到 Pod 已经创建成功了，但 Pod 的状态一会儿是 `Running`，一会儿是 `ExitCode:0`，`READY` 列中始终无法变成 1，而且 `RESTARTS`（重启的数量）的数量不断增加。

通常造成这种现象是因为容器的启动命令不能保持在前台运行。

本例中的 Docker 镜像的启动命令为：

```
zkServer.sh start-background
```

在 Kubernetes 根据 RC 定义创建 Pod 后启动容器，容器的启动命令执行完成时，即认为该容器的运行已经结束，并且是成功结束（`ExitCode=0`）。然后，根据 Pod 的默认重启策略定义（`RestartPolicy=Always`），RC 将启动这个容器。

新的容器执行启动命令后仍然会成功结束，然后 RC 会再次重启该容器，进入一个无限循环的过程中。

解决方法为将 Docker 镜像的启动命令设置为一个前台运行的命令，例如：

```
zkServer.sh start-foreground
```

### 5.2.5 寻求帮助

如果通过系统日志和容器日志都无法找到出现问题的原因，则还可以追踪源码进行分析，或者通过一些在线途径寻求帮助。

- ◎ Kubernetes 的常见问题参见 <https://github.com/GoogleCloudPlatform/kubernetes/wiki/User-FAQ>。
- ◎ Debugging 的常见问题参见 <https://github.com/GoogleCloudPlatform/kubernetes/wiki/Debugging-FAQ>。
- ◎ Service 的常见问题参见 <https://github.com/GoogleCloudPlatform/kubernetes/wiki/Services-FAQ>。
- ◎ StackOverflow 网站关于 Kubernetes 的主题参见 <http://stackoverflow.com/questions/tagged/kubernetes> 或 <http://stackoverflow.com/questions/tagged/google-container-engine>。
- ◎ IRC 频道（#google-containers）参见 <https://botbot.me/freenode/google-containers/>。
- ◎ Kubernetes 邮件列表 Email 参见 [google-containers@googlegroups.com](mailto:google-containers@googlegroups.com)。

## 5.3 Kubernetes 开发中的新功能

本节对 Kubernetes 正在开发中的一些新功能进行介绍，包括 Pod Preset（运行时参数注入策略）、Cluster Federation（集群联邦管理）、CRI（容器运行时接口）、对 GPU 的支持和 Kubernetes 的演进路线等。

### 5.3.1 Pod Preset（运行时参数注入策略）

为了使得 Pod 的定义和运行时参数更加解耦，Kubernetes 从 v1.6 版本开始引入了一个新的资源对象 PodPreset，来定义那些 Pod 启动时所需要的必要信息，在 Pod 启动时进行注入，从而使得 Pod 在定义时尽量减少这些参数的设置。PodPreset 的一个主要设计目标是将 Pod 部署到某个环境时，能够自动化地设置环境变量、用户名密码等应用启动信息，以访问 Kubernetes 集群之外的服务（如数据库服务）。支持的注入参数类型包括环境变量、secret、volume 和 volume mount。

为了使用 PodPreset，首先需要在 API Server 中开启名为 PodPreset 的准入控制器（--admission-control=PodPreset）。PodPreset 作为一个管理对象，与 RC/Service 一样，使用 Label Selector 作用于具有相应 Label 的 Pod。同样，PodPreset 也可以作用于通过 ReplicationController/ ReplicaSet/ Deployment 定义的 Pod，只要 Label Selector 与 Pod 的 Label 可以完成关联。最后，还可以设置多个 PodPreset 作用于一个 Pod。

下面通过几个示例说明 PodPreset 的用法。

### 例 1：使用环境变量设置 PodPreset

以 myweb 应用为例，创建时还无法获知 MySQL 服务的名称和端口号，只需完成 myweb 本身的配置，myweb-pod.yaml 如下：

```
apiVersion: v1
kind: Pod
metadata:
  name: myweb
  labels:
    app: myweb
spec:
  containers:
  - name: myweb
    image: kubeguide/tomcat-app:v1
    ports:
    - containerPort: 8080
```

接下来创建一个 PodPreset，为 myweb Pod 需要访问的数据库服务名和端口号设置相应的环境变量，并通过 Label Selector 设置 myweb 的 Label。myweb-podpreset-db.yaml 的内容如下：

```
kind: PodPreset
apiVersion: settings.k8s.io/v1alpha1
metadata:
  name: myweb-db-setting
spec:
  selector:
    matchLabels:
      app: myweb
  env:
  - name: MYSQL_SERVICE_HOST
    value: 'mysql'
  - name: MYSQL_SERVICE_PORT
    value: '3306'
```

使用 `kubectl create` 命令创建这个 PodPreset：

```
# kubectl create -f myweb-podpreset-db.yaml
podpreset "myweb-db-setting" created
```

接下来创建 myweb Pod：

```
# kubectl create -f myweb-pod.yaml
pod "myweb" created
```

查看这个 Pod 的详细设置，可以看到环境变量已经成功注入 Pod 的定义中，同时 Kubernetes 会在 annotations 中新增一条记录：`podpreset.admission.kubernetes.io/myweb-db-setting: "840497"`，

后面的数字代表 resource version。

```
# kubectl get po/myweb -o yaml
apiVersion: v1
kind: Pod
metadata:
  annotations:
    podpreset.admission.kubernetes.io/myweb-db-setting: "840497"
  labels:
    app: myweb
    name: myweb
    namespace: default
spec:
  containers:
  - env:
    - name: MYSQL_SERVICE_HOST
      value: mysql
    - name: MYSQL_SERVICE_PORT
      value: "3306"
    image: kubeguide/tomcat-app:v1
    name: myweb
    ports:
    - containerPort: 8080
  .....
```

## 例 2：使用 ConfigMap 设置 PodPreset

定义一个 ConfigMap，myweb-env-config.yaml 的内容如下：

```
apiVersion: v1
kind: ConfigMap
metadata:
  name: myweb-env-config
data:
  MYSQL_SERVICE_HOST: "mysql"
  MYSQL_SERVICE_PORT: "3306"
```

定义一个 PodPreset，使用 envFrom 将 ConfigMap 中数据设置为环境变量，myweb-db-setting-cm.yaml 内容如下：

```
kind: PodPreset
apiVersion: settings.k8s.io/v1alpha1
metadata:
  name: myweb-db-setting
spec:
  selector:
    matchLabels:
      app: myweb
```

```
envFrom:
- configMapRef:
  name: myweb-env-config
```

通过 `kubectl create` 命令创建 ConfigMap 和 PodPreset:

```
# kubectl create -f myweb-env-config.yaml
configmap "myweb-env-config" created
# kubectl create -f myweb-db-setting-cm.yaml
podpreset "myweb-db-setting" created
```

再创建 myweb Pod:

```
# kubectl create -f myweb-pod.yaml
pod "myweb" created
```

查看 myweb Pod 的详细设置，可以看到 PodPreset 的设置已经成功注入 Pod 的定义中:

```
# kubectl get po/myweb -o yaml
apiVersion: v1
kind: Pod
metadata:
  annotations:
    podpreset.admission.kubernetes.io/myweb-db-setting: "841932"
  labels:
    app: myweb
    name: myweb
    namespace: default
spec:
  containers:
  - envFrom:
    - configMapRef:
        name: myweb-env-config
      image: kubeguide/tomcat-app:v1
      name: myweb
    ports:
    - containerPort: 8080
      protocol: TCP
```

### 例 3: 多个 PodPreset 作用于一个 Pod

第 1 个 PodPreset 使用例 1 中的定义，myweb-podpreset-db.yaml 的内容如下:

```
kind: PodPreset
apiVersion: settings.k8s.io/v1alpha1
metadata:
  name: myweb-db-setting
spec:
  selector:
    matchLabels:
```

```
    app: myweb
  env:
  - name: MYSQL_SERVICE_HOST
    value: 'mysql'
  - name: MYSQL_SERVICE_PORT
    value: '3306'
```

再定义第 2 个 PodPreset，对 volume 进行设置，myweb-podpreset-volume.yaml 的内容如下：

```
kind: PodPreset
apiVersion: settings.k8s.io/v1alpha1
metadata:
  name: myweb-volume-setting
spec:
  selector:
    matchLabels:
      app: myweb
  volumeMounts:
  - mountPath: /html
    name: webpage-volume
  volumes:
  - name: webpage-volume
    emptyDir: {}
```

通过 `kubectl create` 命令创建 ConfigMap 和 PodPreset：

```
# kubectl create -f myweb-podpreset-db.yaml
podpreset "myweb-db-setting" created
# kubectl create -f myweb-podpreset-volume.yaml
podpreset "myweb-volume-setting" created
```

再创建 myweb Pod：

```
# kubectl create -f myweb-pod.yaml
pod "myweb" created
```

查看 myweb Pod 的详细设置，可以看到两个 PodPreset 的设置都成功注入 Pod 的定义中，同时设置了两条 annotation 标记为通过 PodPreset 进行了 Pod 定义的修改：

```
# kubectl get po/myweb -o yaml
apiVersion: v1
kind: Pod
metadata:
  annotations:
    podpreset.admission.kubernetes.io/myweb-db-setting: "843538"
    podpreset.admission.kubernetes.io/myweb-volume-setting: "843539"
  labels:
    app: myweb
  name: myweb
```



```

  namespace: default
spec:
  containers:
  - env:
    - name: MYSQL_SERVICE_HOST
      value: mysql
    - name: MYSQL_SERVICE_PORT
      value: "3306"
    image: kubeguide/tomcat-app:v1
    name: myweb
    ports:
    - containerPort: 8080
      protocol: TCP
    volumeMounts:
    - mountPath: /html
      name: webpage-volume
  volumes:
  - emptyDir: {}
    name: webpage-volume

```

例 4: PodPreset 与 Pod 中的设置存在冲突时, 保留 Pod 的设置, PodPreset 将不生效

在 Pod 的定义中已存在 volumeMount 的设置, myweb-pod.yaml 的内容如下:

```

apiVersion: v1
kind: Pod
metadata:
  name: myweb
  labels:
    app: myweb
spec:
  containers:
  - name: myweb
    image: kubeguide/tomcat-app:v1
    ports:
    - containerPort: 8080
    volumeMounts:
    - mountPath: /html
      name: webpage-volume
  volumes:
  - emptyDir: {}
    name: webpage-volume

```

定义 PodPreset, 设置与 Pod 不同的 volume, 容器内的挂载目录相同, myweb-podpreset-volume.yaml 的内容如下:

```

kind: PodPreset

```

```
apiVersion: settings.k8s.io/v1alpha1
metadata:
  name: myweb-volume-setting
spec:
  selector:
    matchLabels:
      app: myweb
  volumeMounts:
    - mountPath: /html
      name: other-volume
  volumes:
    - name: other-volume
      emptyDir: {}
```

通过 `kubectl create` 命令创建 `PodPreset` 和 `Pod`:

```
# kubectl create -f myweb-podpreset-volume.yaml
podpreset "myweb-volume-setting" created
# kubectl create -f myweb-pod.yaml
pod "myweb" created
```

查看 `myweb Pod` 的详细设置，只能看到 `Pod` 定义的 `volume` 设置，`PodPreset` 定义的设置不生效。环境变量的设置也是同样的处理方式。

```
# kubectl get po/myweb -o yaml
apiVersion: v1
kind: Pod
metadata:
  labels:
    app: myweb
    name: myweb
    namespace: default
spec:
  containers:
    - image: kubeguide/tomcat-app:v1
      name: myweb
      ports:
        - containerPort: 8080
          protocol: TCP
      volumeMounts:
        - mountPath: /html
          name: webpage-volume
  volumes:
    - emptyDir: {}
      name: webpage-volume
```

`PodPreset` 目前还处于 `Alpha` 开发阶段，有一些使用限制。

(1) PodPreset 的定义需要在创建 Pod 之前准备好，暂时无法在 Pod 运行时动态加载。

(2) PodPreset 对象暂时无法通过 `kubectl get` 命令查看，但可以通过 `kubectl delete` 命令进行删除。

```
# kubectl create -f myweb-podpreset-volume.yaml
podpreset "myweb-db-setting" deleted
```

### 5.3.2 Cluster Federation（集群联邦）

集群联邦从 Kubernetes v1.3 版本开始引入，目标是对多个 Kubernetes 集群进行统一管理，将用户的应用部署到全球各地的不同数据中心或者云环境中，同时通过动态优化部署来节约运营成本。本节介绍 Kubernetes 中 Federation（集群联邦）的主要特性和使用 Federation 管理多集群的原理。

#### 1. Federation 的主要特性

Federation 主要通过以下特性来实现多集群的统一管理。

- ◎ 跨群集资源同步：Federation 提供在多个集群之间保持资源同步的能力，比如通过 Federation 可以确保跨集群的 Deployment 在多个集群中始终同时存在并保持一致。
- ◎ 跨集群服务发现：Federation 提供了自动配置 DNS 服务器和全局负载均衡器（可访问所有 Kubernetes 集群后端服务的负载均衡器）的能力，比如通过 Federation 可以确保使用一条全局虚拟 IP（VIP）或 DNS 记录即可访问部署在多个 Kubernetes 集群中的后端服务。
- ◎ 高可用性：Kubernetes Federation 可以在集群之间分发负载，并且支持自动配置 DNS 服务器和全局负载均衡器，大大降低了发生系统故障的几率，提高了系统的可用性。
- ◎ 避免厂商锁定：Federation 使得应用进行跨不同类型的云平台联合部署变得很容易，而集群中应用程序的迁移也变得更加轻松，因此可以有效地避免出现厂商锁定的情况。

#### 2. Federation 要解决的主要挑战

在 Federation 的实际使用场景中，会面对一些非常重要的挑战。

##### 1) 位置亲和性

在使用多集群部署分布式应用时，前端应用与后端资源（可以 Pod 形式的应用、存储或者其他为前端应用提供服务的资源）的相对位置对于访问延迟、资源开销和系统稳定性具有决定

性的影响。那么 Federation 中应该如何考虑这种位置亲和决策呢？在 Federation 的设计理念中，针对前后端的相对位置，将前后端关系分为三类：严格耦合、严格解耦和优先耦合。三者分别对应前后端必须绑定、可以完全分离及优先绑定这三种应用场景。通过位置亲和性，就可以将严格解耦的应用进行基于 Pod 的平均分配或者随机分配，对优先耦合的应用进行优先分配到同一集群并接受部分移动，而对于严格耦合的应用则严格分配到同一集群环境中。

## 2) 跨集群服务发现

在 Federation 中 Pod 使用外部 DNS 客户端来实现与单集群类似的标准服务发现。DNS 将服务解析为本地集群地址或者外部集群地址。除严格耦合的前后端场景外，前端都可以不用关心 DNS 解析的结果是位于同一集群内还是同一集群外。

## 3) 跨集群应用调度

Federation 的跨集群调度机制中，Federation 控制平面在接收到所有集群的资源对象创建请求后，可以简单地将这个请求重定向给某个集群，也可以将请求“分解”为多个子请求发送给不同的集群。同时，Federation 控制平面需要分析应用的属性（位置亲和性、隐私级别等），并以此作为依据执行更优化的跨集群调度。此外，完善的跨集群调度机制还需要支持准入控制机制、自动扩容和缩容机制、故障重调度机制及基于计算能力的调度优化等。

## 4) 跨集群应用迁移

在 Federation 的使用过程中，可能会遇到部分集群容量将满、转换云供应商、变换核心集群位置等需要进行应用迁移的场景。在这种情况下，Federation 的跨集群迁移工作是按照应用位置亲和性来分别进行的：对严格解耦的应用采取一次或多次分步迁移的方式进行，每次迁移的粒度也很自由；对优先耦合的应用，需要首先找到具有足够多的资源容量可以容纳待迁移应用的目标集群，并锁定该部分的资源容量，之后按照特定的顺序在特定的时间内完成迁移工作；而对于严格耦合的程序而言，除了需要符合与优先耦合类似的资源要求，在迁移过程中还需要考虑是否能满足数据一致性和应用一致性的要求，如果不能满足要求，则不建议直接进行迁移。

## 5) 故障隔离

Federation 保留了 Kubernetes 集群的应用隔离机制，一般情况下并不会显著地增加多个集群之间故障的关联性。Federation 控制平面与每个 Kubernetes 集群的控制平面是严格独立的，Federation 控制平面的故障应不影响每个 Kubernetes 自身的正常运行。

- ◎ 统一监控、统一预警和跨集群联合审计。
- ◎ 统一认证授权、跨集群的配额管理。

### 3. Federation 的架构设计

针对这些调整，Federation 的整体架构设计采用了解耦和分层的思路：Federation 控制平面位于所有 Kubernetes 集群之上，而每个 Kubernetes 集群都是可以独立运行的。除了部分基础配置信息，每个 Kubernetes 集群都不知道其他 Kubernetes 集群的存在，也不知道 Federation 控制平面的存在。在这种设计中，Federation 控制平面就像每个 Kubernetes 集群的 API 客户端，因此与每个集群解除了耦合关系。与 Federation 解耦和分层的架构相对的是一体式架构设计：即为每个 Kubernetes 集群搭建一个控制平面，这个控制平面只负责管理这个 Kubernetes 集群，而多个控制平面之间通过通信的方式来实现对所有 Kubernetes 集群的联合管理。

Federation 的主要架构如图 5.25 所示。

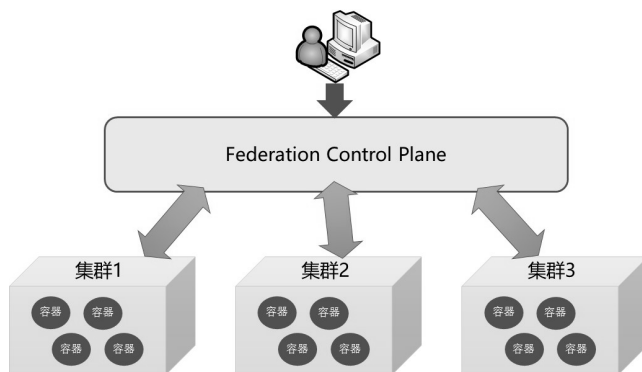


图 5.25 Federation 的主要架构

Federation 分层解耦的设计具有如下优势。

#### 1) 故障隔离性好

如前面所述，Federation 分层解耦的设计保证了 Federation 控制平面与集群的隔离，每个集群和 Federation 控制平面都可以独立运行，出现故障时可以进行单独隔离而互不影响。另外，每个集群和 Federation 控制平台的软件和配置都可以进行独立更新，这为系统的维护提供了极大的便利。

#### 2) 失效几率更低

整体而言，分层设计的系统比一体式设计的系统出现故障的概率要高（概率叠加），但由于各系统解除了耦合，所以系统完全失效的几率要低于一体式设计的系统。

#### 3) 可扩展性高

在 Federation 的分层解耦设计中，每个底层的 Kubernetes 集群内部都可以完全独立地进行扩展，而 Federation 中也可以很容易扩展加入新的集群而不影响现有集群。基于分层架构

的优势，在未来的 Kubernetes 版本里，甚至可能会提供集群联邦的联邦功能（Federation of Federation）。

#### 4) 代码模块化和分离

Federation 控制平面的代码与每个云供应商的 Kubernetes 控制平面的代码是分离的，它们之间通过共享库的方式来实现部分代码共享。这种设计允许 Kubernetes 和 Federation 的代码开发工作在很大程度上独立进行，同时促进了更好的代码模块化和独立的接口设计。

#### 5) 更灵活的管理策略

一体式设计的系统看似管理工作更简单，但是由于不同的云供应商和本地数据中心有不同的特点（硬件、定价、限制等），而 Federation 的分层解耦架构允许独立地管理每个 Kubernetes 集群，这虽然看似提高了管理的复杂性，但是对于整个系统而言，提供了更丰富的控制手段和管理策略。

- ◎ 更好的代码模块化和界面设计。
- ◎ 控制平面的成本更低。

在一体式设计的系统中，每个 Kubernetes 集群均需要部署自己的控制平面，而在分层解耦的设计中，Federation 的控制平面只需要独立部署一次。如果我们需要实现控制平面的高可用，那么也只需要再部署一个 Federation 控制平面。可见 Federation 的分层设计在控制平面的成本开销上的优势非常明显。

### 4. Federation 的优势

Federation 是 Kubernetes 多集群的解决方案，如果不需要使用多个 Kubernetes 集群，则 Federation 并没有太大用处。在下列情况下，可以考虑引入 Federation。

- ◎ 低延迟：通过多地区部署服务，配合就近选择集群提供服务的方式，Federation 可以最小化服务访问的延迟。
- ◎ 故障隔离：当系统发生故障时，由多个小型集群（这些集群通常分布在不同的区域）组成的 Federation 比单个大型集群更适合快速有效地隔离，而且对整体服务的影响会更小。
- ◎ 可扩展性：根据谷歌的经验，在超大规模的生产环境中，单个 Kubernetes 集群的扩展性受到集群规模的制约。当单个集群规模过大时，集群性能下降。而多集群 Federation 则可以提高集群规模的上限，提供更好的可扩展性。
- ◎ 混合云：Federation 支持私有云和公有云的组合，可以使用 Federation 在不同的云供应商或多个本地数据中心上搭建多个 Kubernetes 集群，实现混合云部署。

## 5. Federation 的局限性

除了上述优势，在使用 Federation 之前也应充分考虑一些潜在问题。

- ◎ 网络带宽和成本的增加：为确保所有的集群运行状态符合预期，Federation 控制平面会持续监控所有集群。如果 Federation 中的集群运行在同一个云供应商的不同地区上（一般同一云供应商跨地区的网络通信是需要收费的）或者运行在不同的云供应商上，那么将会导致显著的网络开销和成本的提升。
- ◎ 削弱了多集群之间的隔离性：Federation 控制平面一旦出现问题，就可能会影响到所有的集群。一种可能的方案是，通过尽可能减少 Federation 控制平面中的逻辑，将 Federation 控制平面的逻辑尽可能多地传递给各 Kubernetes 子集群的控制平面，来减缓这种情况。但这类问题很难完全避免。同样，目前 Federation 这种“中心控制”的设计思路和实现方式还可能导致安全性及多集群同时不可用方面的问题。
- ◎ 成熟度：相对而言，Federation 出现较晚，还不是很成熟。目前 Kubernetes 中的资源对象只有一部分在 Federation 中是可用的，而且很多可用的资源对象目前还只是 Alpha 状态。此外，Federation 的设计、实现和用法目前随着 Kubernetes 大版本的变更都发生了很多改变，因此给使用者带来了不少困难。

### 5.3.3 容器运行时接口（Container Runtime Interface-CRI）

归根结底，Kubernetes Node（kubelet）的主要功能就是启动和停止容器的组件，我们称之为容器运行时（Container Runtime），其中最知名的就是 Docker 了。为了让 Kubernetes 更具扩展性，从其 v1.5 版本开始加入了容器运行时插件 API，我们称之为 Container Runtime Interface，简称 CRI。

#### 1. CRI 概述

每种容器运行时都有其特点，因此不少用户希望 Kubernetes 能够支持更多的运行时。Kubernetes 从 v1.5 版本开始引入了 CRI 接口规范，通过插件接口模式，让 Kubernetes 无须重新编译就可以使用更多的容器运行时。CRI 包含 Protocol Buffers、gRPC API、运行库支持及开发中的标准规范和工具。Docker-CRI 的实现在 Kubernetes v1.6 版本时更新为 Beta 版本，并在 kubelet 启动时默认启动。

可替代的容器运行时支持是 Kubernetes 中的新概念。在 Kubernetes v1.3 发布时，rkt 项目同时发布，让 rkt 容器引擎成为除 Docker 外的又一选择。然而，不管是 Docker 还是 rkt，都用到了 kubelet 的内部接口，同 kubelet 源码纠缠不清。这种程度的集成需要对 kubelet 的内部

机制有非常深入的了解，还会给社区带来管理压力，这就给新生代容器运行时造成了难于跨越的集成壁垒。CRI 接口规范试图用定义清晰的抽象层清除这一壁垒，让开发者能够专注于容器运行时本身。在通向插件式容器支持及建设健康生态环境的路上，这是一小步，也是重要的一步。

## 2. CRI 的主要组件

kubelet 使用 gRPC 框架利用 UNIX Socket 与容器运行时（或 CRI 代理）进行通信。在这个过程中 kubelet 是客户端，CRI 代理（shim）是服务端，如图 5.26 所示。

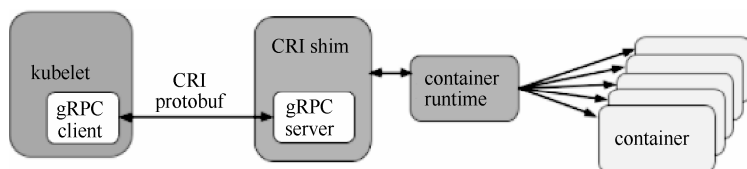


图 5.26 CRI 的主要组件

Protocol Buffers API 包含两个 gRPC 服务：ImageService 和 RuntimeService。

ImageService 提供了从仓库拉取镜像、查看和移除镜像的功能。

RuntimeService 负责 Pod 和容器的生命周期管理、与容器的交互（exec/attach/port-forward）。rkt 和 Docker 这样的容器运行时可以使用一个 Socket 同时提供两个服务。在 kubelet 中可以用 --container-runtime-endpoint 和 --image-service-endpoint 参数设置这个 Socket。

## 3. Pod 和容器的生命周期管理

Pod 由一组应用容器组成，其中包含了共有的环境和资源约束。在 CRI 里，这个环境被称为 PodSandbox。Kubernetes 有意为容器运行时留下一些发挥空间，它们可以根据自己的内部实现来解释 PodSandbox。对于 Hypervisor 类的运行时，PodSandbox 会具体化为一个虚拟机。其他的例如 Docker，会是一个 Linux 命名空间。在 v1alpha1 API 中，kubelet 会创建 Pod 级别的 cgroup 传递给容器运行时，并以此运行所有进程来保障 PodSandbox 对 Pod 的资源保障。

在启动 Pod 之前，kubelet 调用 RuntimeService.RunPodSandbox 来创建环境。这一过程包括为 Pod 设置网络资源（分配 IP 等操作）。PodSandbox 激活之后，就可以独立地创建、启动、停止和删除不同的容器了。kubelet 会在停止和删除 PodSandbox 之前首先停止和删除其中的容器。

kubelet 的职责在于通过 RPC 管理容器的生命周期，实现容器生命周期的钩子，以及存活和健康监测，执行 Pod 的重启策略等。



RuntimeService 服务包括对 Sandbox 和 Container 操作的方法，下面的伪代码展示了主要的 RPC 方法：

```
service RuntimeService {
    // 沙箱操作
    rpc RunPodSandbox(RunPodSandboxRequest) returns (RunPodSandboxResponse) {}
    rpc StopPodSandbox(StopPodSandboxRequest) returns (StopPodSandboxResponse) {}
    {}
    rpc RemovePodSandbox(RemovePodSandboxRequest) returns
(RemovePodSandboxResponse) {}
    rpc PodSandboxStatus(PodSandboxStatusRequest) returns
(PodSandboxStatusResponse) {}
    rpc ListPodSandbox(ListPodSandboxRequest) returns (ListPodSandboxResponse)
    {}
    // 容器操作
    rpc CreateContainer(CreateContainerRequest) returns
(CreateContainerResponse) {}
    rpc StartContainer(StartContainerRequest) returns (StartContainerResponse)
    {}
    rpc StopContainer(StopContainerRequest) returns (StopContainerResponse) {}
    rpc RemoveContainer(RemoveContainerRequest) returns
(RemoveContainerResponse) {}
    rpc ListContainers(ListContainersRequest) returns (ListContainersResponse)
    {}
    rpc ContainerStatus(ContainerStatusRequest) returns
(ContainerStatusResponse) {}
    ...
}
```

#### 4. CRI 设计为面向容器而不是 Pod

众所周知，Kubernetes 的最小调度单元是 Pod。曾经可能采用的一个 CRI 设计就是复用 Pod 对象，使得容器运行时可以自行实现控制逻辑和状态转换，这样一来，就能极大地简化 API，让 CRI 能够更广泛地适用于多种容器运行时。但是经过深入讨论之后，Kubernetes 放弃了这一想法。

首先，kubelet 有很多 Pod 级别的功能和机制（例如 crash-loop backoff 机制），如果交给容器运行时去实现，则会造成很重的负担；其次，更重要的是，Pod 标准还在快速演进。很多新功能（例如初始化容器）是由 kubelet 直接管理容器的，而无须交给容器运行时实现。

CRI 选择了在容器级别进行实现，使得容器运行时能够共享这些通用特性，以获得更快的开发速度。这并不意味着设计哲学的改变——kubelet 要负责保证容器应用实际状态和声明状态的一致性。

Kubernetes 为用户提供了和 Pod 及其中的容器进行交互的能力（`kubectl exec/attach/port-forward`）。kubelet 目前提供了两种方式来支持这些功能。

（1）调用容器的本地方法。

（2）使用 Node 上的工具（例如 `nsenter` 及 `socat`）。

因为多数工具假设 Pod 用 Linux namespace 做了隔离，因此使用 Node 上的工具并不是一个容易移植的方案。在 CRI 中显式定义这些调用方法，让容器运行时进行具体实现。下面的伪代码显示 `exec/attach/port-forward` 这几个调用需要实现的 `RuntimeService` 方法：

```
service RuntimeService {
    ...
    // ExecSync 在容器中同步执行一个命令。
    rpc ExecSync(ExecSyncRequest) returns (ExecSyncResponse) {}
    // Exec 在容器中执行命令
    rpc Exec(ExecRequest) returns (ExecResponse) {}
    // Attach 附着在容器上
    rpc Attach(AttachRequest) returns (AttachResponse) {}
    // PortForward 从 Pod 沙箱中进行端口转发
    rpc PortForward(PortForwardRequest) returns (PortForwardResponse) {}
    ...
}
```

目前还有一个潜在问题是，kubelet 处理所有的请求连接，使其有成为 Node 通信瓶颈的可能。在设计 CRI 时，让容器运行时能够跳过中间过程。容器运行时可以启动一个单独的流式服务来处理请求（还能对 Pod 的资源使用进行记录），并将服务地址返回给 kubelet。这样 kubelet 就能反馈信息给 API Server，使之可以直接连接到容器运行时提供的服务，并连接到客户端。

## 5. 尝试使用新的 Docker-CRI 来创建容器

要尝试新的 Kubelet-CRI-Docker 集成，则只需为 kubelet 启动参数加上 `--enable-cri=true` 开关来启动 CRI。这个选项从 Kubernetes v1.6 开始已经作为 kubelet 的默认选项了。如果不希望使用 CRI，则可以设置 `--enable-cri=false` 来关闭这个功能。

查看 kubelet 的日志，可以看到启用 CRI 和创建 gRPC Server 的日志：

```
I0603 15:08:28.953332 3442 container_manager_linux.go:250] Creating
Container Manager object based on Node Config: {RuntimeCgroupsName: SystemCgroupsName:
KubeletCgroupsName: ContainerRuntime:docker CgroupsPerQOS:true CgroupRoot:/
CgroupDriver:cgroupfs ProtectKernelDefaults:false EnableCRI:true
NodeAllocatableConfig:{KubeReservedCgroupName: SystemReservedCgroupName:
EnforceNodeAllocatable:map[pods:{}] KubeReserved:map[] SystemReserved:map[]
HardEvictionThresholds:[{Signal:memory.available Operator:LessThan
Value:{Quantity:100Mi Percentage:0} GracePeriod:0s MinReclaim:<nil>}}}
```

```
ExperimentalQOSReserved:map[]}
.....
I0603 15:08:29.060283    3442 kubelet.go:573] Starting the GRPC server for the
docker CRI shim.
```

创建一个 Deployment:

```
$ kubectl run nginx --image=nginx
deployment "nginx " created
```

查看 Pod 的详细信息，可以看到将会创建 Sandbox 的 Event:

```
$ kubectl describe pod nginx
.....
Events:
...From                                Type      Reason                                Message
...-----
...default-scheduler                   Normal    Scheduled                             Successfully assigned nginx to
k8s-node-1
...kubelet, k8s-node-1                 Normal    SandboxReceived                       Pod sandbox received, it will
be created.
.....
```

这表明 kubelet 使用了 CRI 接口来创建容器。

## 6. CRI 的进展

虽然 CRI 还比较初级，但也已经有了很多项目在尝试把各种容器运行时纳入 CRI。

- ◎ cri-o: OCI 兼容运行时。
- ◎ rktlet: rkt 容器运行时。
- ◎ frakti: 基于 Hypervisor 的容器运行时。
- ◎ Docker CRI 代理。

如果有兴趣集成新的容器运行时，则可以参阅 CRI 的开发者指南，获取 API 中已知的限制和问题，通过参与改进有助于项目的成长。

### 5.3.4 对 GPU 的支持

随着人工智能和机器学习的迅速发展，GPU 计算变得越来越重要。Kubernetes 的发展规划中 GPU 资源占有非常重要的地位。用户应该能够为他们的工作任务请求 GPU 资源，就像请求 CPU 或内存一样，Kubernetes 将调度容器到具有 GPU 资源的节点上去。目前 Kubernetes 对 NVIDIA GPU 进行了实验性的支持。

下面对 Kubernetes 如何管理容器请求 GPU 资源进行说明。

## 1. 前提条件

(1) Kubernetes 工作节点必须已预先安装好 NVIDIA 驱动程序，否则 kubelet 将无法检测到 NVIDIA GPU。如果 kubelet 还是没能成功将 NVIDIA GPU 设置为节点资源的一部分，则可以尝试通过重装节点 Nvidia 驱动来解决。

(2) 由于是实验性特性，因此需要为每个 Kubernetes 工作节点的 kubelet 开启该特性，即在每个节点 kubelet 的启动参数中加上：--feature-gates="Accelerators=true"。

(3) 目前必须使用 Docker 作为容器引擎才能使用 GPU，暂不支持其他容器引擎。

在完成上述配置后，Kubernetes 节点将自动发现并设置所有的 NVIDIA GPU 作为集群中的可调度资源。

## 2. 容器 GPU 资源请求的配置

NVIDIA GPU 在 Kubernetes 中的资源名称为 alpha.kubernetes.io/nvidia-gpu，可以对其进行容器级别的 GPU 资源请求设置。下面的配置为两个容器分别设置了 GPU 资源请求。

```
apiVersion: v1
kind: pod
spec:
  containers:
    - name: gpu-container-1
      resources:
        limits:
          alpha.kubernetes.io/nvidia-gpu: 2 # 需要两个 GPU
    - name: gpu-container-2
      resources:
        limits:
          alpha.kubernetes.io/nvidia-gpu: 3 # 需要 3 个 GPU
```

由于是实验特性，因此目前对 GPU 的资源配置有很多限制。

- ◎ GPU 请求目前只能在 limits 字段进行设置，还不支持 requests 字段的设置。
- ◎ 容器与容器之间不能共享 GPU，Pod 之间也不能共享。
- ◎ 每个容器只能请求整数个（一个或多个）GPU，不能请求单个 GPU 的一部分。
- ◎ 集群工作节点预期是同质的，也就是说运行着相同的 GPU 硬件。

如果集群中运行着不同版本的 GPU，那么我们可以通过使用 Node Label（节点标签）和 Node Selector（节点选择器）将 Pod 调度到合适的 GPU 所属节点。以下是这个工作流程的说明。

(1) 首先在节点上添加 kubelet 启动参数--node-labels，设置一个 Label “alpha.kubernetes.io/nvidia-gpu-name” 为 GPU 硬件类型的信息：

```
# NVIDIA_GPU_NAME=$(nvidia-smi --query-gpu=gpu_name --format=csv,noheader
--id=0)
# source /etc/default/kubelet
# KUBELET_OPTS="$KUBELET_OPTS
--node-labels='alpha.kubernetes.io/nvidia-gpu-name=$NVIDIA_GPU_NAME'"
# echo "KUBELET_OPTS=$KUBELET_OPTS" > /etc/default/kubelet
```

(2) 然后，使用节点亲和性规则来指定 Pod 可以被调度到的 Node：

```
kind: pod
apiVersion: v1
metadata:
  annotations:
    scheduler.alpha.kubernetes.io/affinity: >
    {
      "nodeAffinity": {
        "requiredDuringSchedulingIgnoredDuringExecution": {
          "nodeSelectorTerms": [
            {
              "matchExpressions": [
                {
                  "key": "alpha.kubernetes.io/nvidia-gpu-name",
                  "operator": "In",
                  "values": ["Tesla K80", "Tesla P100"]
                }
              ]
            }
          ]
        }
      }
    }
spec:
  containers:
    - name: gpu-container-1
      resources:
        limits:
          alpha.kubernetes.io/nvidia-gpu: 2
```

上面的配置确保 Pod 将被调度到含有 Label “alpha.kubernetes.io/nvidia-gpu-name” 且值为 “Tesla K80” 或 “Tesla P100” 的节点上。

### 3. 容器对 NVIDIA CUDA 库的访问

CUDA (Compute Unified Device Architecture) 是由 NVIDIA 公司推出的通用并行计算平台，

包含了 CUDA 指令集架构（ISA）及 GPU 内部的并行计算引擎，非常适合需要大规模并行计算的领域，例如图形动画、科学计算、地质、生物、物理模拟等。

假设各 Node 已安装好 CUDA 库，则 Node 上应该存在 `/usr/lib/nvidia-367` 目录，可以通过挂载 `hostPath` 存储卷的方式，让 Pod 访问 CUDA 库文件。一个需要访问 NVIDIA CUDA 库的 Pod 的配置样例如下：

```
kind: Pod
apiVersion: v1
metadata:
  name: gpu-pod
spec:
  containers:
  - name: gpu-container-1
    securityContext:
      privileged: true
    resources:
      limits:
        alpha.kubernetes.io/nvidia-gpu: 1
    volumeMounts:
      - mountPath: /usr/local/nvidia/bin
        name: bin
      - mountPath: /usr/lib/nvidia
        name: lib
  volumes:
  - hostPath:
      path: /usr/lib/nvidia-367/bin
      name: bin
  - hostPath:
      path: /usr/lib/nvidia-367
      name: lib
```

#### 4. 未来发展

- ◎ Kubernetes 目前已经实现了对硬件加速器的支持，不过仍处于早期阶段，后面会有很多更新。
- ◎ GPU 和其他加速器将很快成为 Kubernetes 系统的原生计算资源类型，而无须诸多额外的配置工作。
- ◎ 目前的 API 限制还有很多，未来会推出功能更丰富的 API，能支持以可扩展的形式进行 GPU 等硬件加速器资源的供给、调度和使用。
- ◎ Kubernetes 将能自动确保使用 GPU 的应用程序得到最佳的性能表现。
- ◎ 访问 CUDA 库等关键问题将被 Kubernetes 原生方案很好地解决。

### 5.3.5 Kubernetes 的演进路线（Roadmap）和开发模式

Kubernetes 将每个版本的待开发功能和未完成的功能在 [kubernetes/features](https://github.com/kubernetes/features) 网页统一发布和管理，网址为 <https://github.com/kubernetes/features>，目前可以跟踪从 v1.3 到 v1.7 版本的功能列表，如图 5.27 所示。

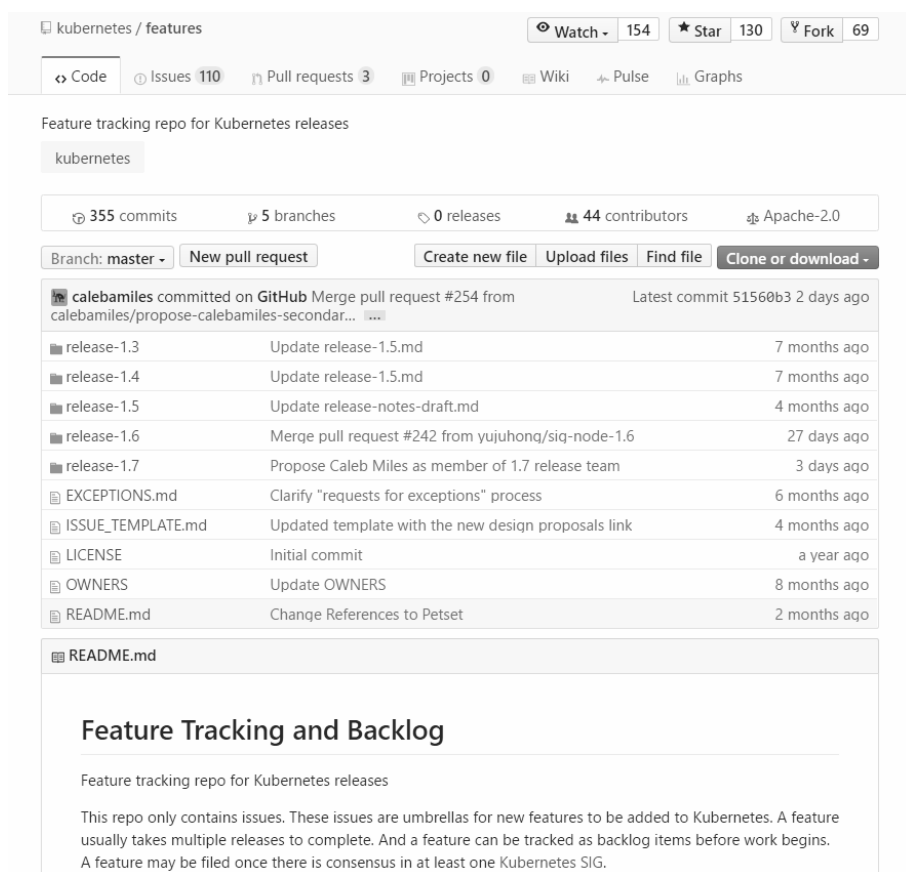


图 5.27 Kubernetes 功能特性网页

以 release-1.6 为例，从 release-1.6.md 文档中可以查看这个大版本的开发计划，如图 5.28 所示。

根据文中的链接“Feature tracking spreadsheet link”，可以查看到该版本所包含的全部功能列表，按开发阶段分为 Alpha、Beta 和 Stable 三个类别，可以很直观地看到各功能模块的实现阶段。单击第 1 列的链接还可以打开相关功能的 issue 页面，进一步查看该功能的详细信息，如图 5.29 所示。

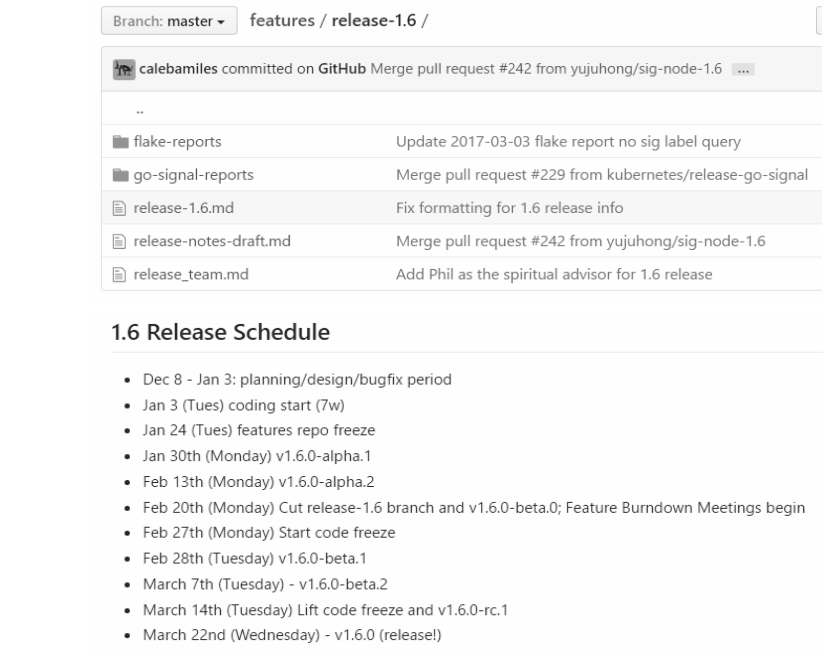


图 5.28 大版本开发计划

	A	B	C	D	E
1	Taints/tolerations including forgiveness (tolerationSeconds)	SIG	Assignee	One Line Release Note Description	
2	Alpha	8			
3	Beta	12			
4	Stable	9			
5	TOTAL	29			
6	Monitoring Pipeline Metrics HPA API	Alpha SIG-Autoscaling	DirectXMan12	The Horizontal Pod Autoscaler now supports drawing metrics through the API server aggregator.	
7	Arbitrary Custom Metrics in the Horizontal Pod Autoscaler	Alpha SIG-Autoscaling	DirectXMan12	The Horizontal Pod Autoscaler now supports scaling on multiple, custom metrics.	
8	Integrate Cluster Bootstrap/Discovery with Kubernetes Core	Alpha SIG-ClusterLifecy	jbeda	New Bootstrap Token authentication and management method. Works well with kubeadm.	
9	Support out-of-process and out-of-tree cloud providers	Alpha SIG-ClusterLifecy	wliao0	Adds a new cloud-controller-manager binary that may be used for testing the new out-of-core cloudprovider flow.	
10	Represent node problems using NoExecute taint, and allow user-defined	Alpha SIG-Scheduling	gnanok	Per-pod configurable duration to stay bound to a node that becomes unreachable, not ready, or has other problems.	
11	Pod Injection Policy	Alpha SIG-Service-Catal	ysroftaz	Adds a new API resource "PodPrevent" and admission controller to enable defining cross-cutting injection of Vol	
12	All in one volume proposal	Alpha SIG-Storage	peeler	A new volume driver capable of projecting secrets, configmaps, and downward API items into the same directo	
13	Flex volume API and Improved Lifecycle (flexvolume)	Alpha SIG-Storage	chukri-nelluri	Flex volume plugin is updated to support attach/detach interfaces. It broke backward compatibility. Please updat	
14	DaemonSet updates	Beta SIG-Apps	lukaszoz	Implement the update feature for DaemonSet.	
15	Allow deployments to correctly indicate they are failing to deploy	Beta SIG-Apps	kargakis	Deployments that cannot make progress on rolling out the newest version will now indicate via the API that they are	
16	Role-based access control	Beta SIG-Auth	lgitt	RBAC API is promoted to v1beta1 (rbac.authorization.k8s.io/v1beta1), and defines default roles for control plan	
17	Kubelet TLS Bootstrap	Beta SIG-ClusterLifecy	gtank	Introduces an API for the Kubelet to request TLS certificates from the API server.	
18	Dramatically Simplify Kubernetes Cluster Creation	Beta SIG-ClusterLifecy	jbeda	kubeadm is enhanced and improved with a baseline feature set and command line flags that are now marked as	
19	It should be fast and painless to deploy a Federation of Kubernetes clus	Beta SIG-Federation	madhusudanc	"kubefed" has graduated to beta, supports hosting federation on on-prem clusters, automatically configures "ku	
20	Redefine the Container Runtime Interface	Beta SIG-Node	yujuhong	The Docker CRI implementation is Beta and is enabled by default in kubelet. You can disable it by --enable-cr	
21	Inter-pod affinity/anti-affinity	Beta SIG-Scheduling	davidopp	Rules for spreading and packing pods relative to one another, within arbitrary topologies (node, zone, etc.)	
22	Node affinity	Beta SIG-Scheduling	davidopp	Rules for restricting which node(s) a pod can schedule onto	
23	Multiple user-defined schedulers	Beta SIG-Scheduling	linofthysc	User can run multiple schedulers in parallel responsible for different sets of pods	
24	Taints/tolerations including forgiveness (tolerationSeconds)	Beta SIG-Scheduling	davidopp	Rules for "tainting" pods from nodes by default (support use cases like dedicated nodes, and reserve nodes v	
25	Support Volume Mount Options	Beta SIG-Storage	grufted	Added support for mount options in persistent volumes.	
26	etcd v3 as storage backend for APIServer	Stable SIG-API-machine	timothy-sc	The internal storage layer for Kubernetes cluster state has been updated to use etcd v3 by default for new clus	
27	Add support for pod and namespace groups	Stable SIG-Node	denekowynec	Kubelet launches pods in a new group hierarchy to better enforce quality of service. Operators must drain all	
28	Configurable Dynamic Provisioning aka StorageClass	Stable SIG-Storage	jsafrane	StorageClass API is promoted to v1 (storage.k8s.io/v1)	
29	Default Storage Classes for Cloud Providers	Stable SIG-Storage	jsafrane	Default storage classes are deployed during installation on Azure, AWS, GCE, OpenStack and vSphere	
30	Create environment variables from all keys in a Secret/Configmap	Stable SIG-Storage	pmorie	Populate environment variables from a configmap or secret.	
31	External provisioners	Stable SIG-Storage	wongma7	Support for user-written/run dynamic PV provisioners. See github.com/kubernetes-incubator/external-storage f	
32	Default CMC, ScaleIO Volume Plugin	Stable SIG-Storage	vladimirv	ScaleIO Kubernetes Volume Plugin added enabling pods to seamlessly access and use data stored on ScaleIO	
33	Portworx Volume Plugin	Stable SIG-Storage	eddyedev	Portworx Volume Plugin added capability to use (Portworx)http://www.portworx.com) as a storage provider for	
34	Customized mounts in chroot to GA	Stable SIG-Storage	jngui97	Add support to use NFSv3, NFSv4, and GlusterFS on GCI image cluster	

图 5.29 特性跟踪表格

每个大版本中的各个 Feature 都由一个特别兴趣小组（Special Interest Group, SIG）负责，SIG 的介绍在网页 <https://github.com/kubernetes/community/blob/master/sig-list.md> 可以找到，如图 5.30 所示。



43 lines (35 sloc) | 11 KB

Raw Blame History

## SIGs and Working Groups

Most community activity is organized into Special Interest Groups (SIGs), time bounded Working Groups, and the community meeting.

SIGs follow these guidelines although each of these groups may operate a little differently depending on their needs and workflow.

Each group's material is in its subdirectory in this project.

When the need arises, a new SIG can be created

### Master SIG List

Name	Leads	Group	Slack Channel	Meetings
API Machinery	@lavalamp Daniel Smith, Google @deads2k David Eads, Red Hat	Group	#sig-api-machinery	Every other Wednesday at 11:00 AM PST
Apps	@michelleN (Michelle Noorali, Deis) @mattfarina (Matt Farina, HPE)	Group	#sig-apps	Mondays 9:00AM PST
Auth	@ericchiang (Eric Chiang, CoreOS) @liggitt (Jordan Liggitt, Red Hat) @deads2k (David Eads, Red Hat)	Group	#sig-auth	Biweekly Wednesdays at 1100 to 1200 PT
Autoscaling	@fgrzadkowski (Filip Grzadkowski, Google) @directxman12 (Solly Ross, Red Hat)	Group	#sig-autoscaling	Biweekly (or triweekly) on Thurs at 0830 PT
AWS	@justinsb (Justin Santa Barbara) @kris-nova (Kris Nova) @chrisslovecnm (Chris Love)	Group	#sig-aws	We meet on Zoom, and the calls are scheduled via the official group mailing list

图 5.30 特别兴趣小组 SIG

目前已经成立的 SIG 小组有 25 个，涵盖了安全、自动扩容和缩容、大数据、AWS 云、文档、网络、存储、调度、UI、Windows 等方方面面，为完善 Kubernetes 的功能群策群力，共同开发。有兴趣、有能力的读者可以申请加入感兴趣的 SIG 小组，并可以通过 Slack 聊天频道与来自世界各地的开发组成员开展技术探讨和解决问题。同时，可以参加 SIG 小组的周例会，共同参与一个功能模块的开发工作。

# 第 6 章

## Kubernetes 源码导读

---

### 6.1 Kubernetes 源码结构和编译步骤

---

Kubernetes 的源码现在托管在 GitHub 上，地址为 <https://github.com/googlecloudplatform/kubernetes>。

编译脚本存放在 `build` 子目录下，在 Linux 环境（可以是虚拟机）中执行如下命令即可完成代码的编译过程：

```
git clone https://github.com/GoogleCloudPlatform/kubernetes.git
cd kubernetes/build
./release.sh
```

制作 `release` 的过程其实有不少有意思的事情发生，包括启动 Docker 容器来安装 Go 语言环境、`etcd` 等，读者若有兴趣则可以查看 `release.sh` 脚本。另外，如果编译环境是通过 HTTP 代理上网的，则需要设置好 Git 与 Docker 相关的 HTTP 代理参数，同时在文件 `kubernetes/build/build-image/Dockerfile` 中增加如下 HTTP 代理参数。

- ◎ `ENV http_proxy=http://username:password@proxyaddr:proxyport。`
- ◎ `ENV https_proxy=http://username:password@proxyaddr:proxyport。`

在编译过程中产生的与 Docker 相关的 `docker image`、`dockerfile` 及编译好的二进制文件包，则存放在 `kubernetes/_output` 目录下，这个目录总共有 4 个子目录：`dockerized`、`images`、`release-stage`、`release-tars`，我们关心后两个目录，其中 `release-stage` 目录下存放的是支持 `linux-amd64` 架构的 Server 端的二进制可执行文件（放在 `server` 子目录下），以及支持不同平台的 Client 端的二进制可执行文件（放在 `client` 子目录下），`release-tars` 则存放的是 `release-stage` 目录下各级子目录的压缩包，与从官方网站下载的完全一样。

考虑到学习和调试 Kubernetes 代码的便利性，我们接下来介绍如何在 Windows 的 LiteIDE 开发环境中完成 Kubernetes 代码的编译和调试。本文假设 Windows 上的 GO 运行框架和 LiteIDE 开发环境已经建立好，并通过 `git clone` 命令已经将 `https://github.com/GoogleCloudPlatform/kubernetes.git` 下载到本地 `C:\kubernetes` 目录中。通过分析 Kubernetes 的目录结构，我们发现 Kubernetes 的源码都在 `pkg` 子目录下。接下来建立 `k8s` 工程目录，目录位置为 `C:\project\go\k8s`，并在里面建立 `src`、`pkg` 两个子目录，然后把 `C:\kubernetes\Godeps\workspace\src` 全部转移到 `C:\project\go\k8s\src` 目录下，因为这里是 Kubernetes 源码的所有依赖包，所以如果手动一个一个地下载，则恐怕以国内的网速一天也搞不定。转移完成后，`C:\project\go\k8s\src` 的目录结构包括如下内容：

```
C:\project\go\k8s\src>dir
2015-07-14 11:56 <DIR>      bitbucket.org
2015-07-14 11:56 <DIR>      code.google.com
2015-07-17 12:30 <DIR>      github.com
2015-07-14 11:56 <DIR>      golang.org
2015-07-14 11:56 <DIR>      google.golang.org
2015-07-14 11:56 <DIR>      gopkg.in
2015-07-14 11:56 <DIR>      speter.net
```

接下来把 `C:\kubernetes` 的整个目录移动到 `C:\project\go\k8s\src\github.com\GoogleCloudPlatform\` 下，因为 Kubernetes 的源码包的完整名字为 “`github.com/GoogleCloudPlatform/kubernetes/pkg`”。上述工作完成以后，所有的源码都在 `C:\project\go\k8s\src` 目录下了，我们用 LiteIDE 打开 `C:\project\go\k8s`，单击菜单 “查看” → “管理 Gopath” → 添加目录 “`C:\project\go\k8s`”，然后可以进入目录 `github.com/GoogleCloudPlatform/kubernetes/pkg` 下，逐一编译每个 `package` 目录了，如图 6.1 所示。

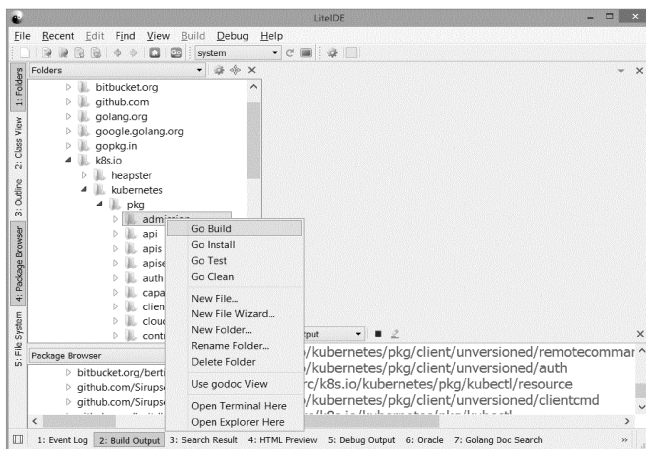


图 6.1 LiteIDE 编译 Kubernetes 的 package

在每个 package 都编译完成以后，我们可以尝试启动 kube-scheduler 进程：在 LiteIDE 里打开 [github.com/GoogleCloudPlatform/kubernetes/pkg/plugin/cmd/kube-scheduler/scheduler.go](https://github.com/GoogleCloudPlatform/kubernetes/pkg/plugin/cmd/kube-scheduler/scheduler.go)，并且按快捷键 Ctrl+R，你会惊奇地发现这个 Kubernetes 服务器端进程竟然也能在 Windows 下运行起来。以下是 LiteIDE 输出的控制台日志：

```
c:/go/bin/go.exe build -i [C:/project/go/k8s/src/github.com/GoogleCloudPlatform/
kubernetes/plugin/cmd/kube-scheduler]
成功：进程退出代码 0。
C:/project/go/k8s/src/github.com/GoogleCloudPlatform/kubernetes/plugin/cmd/
kube-scheduler/kube-scheduler.exe [C:/project/go/k8s/src/github.com/GoogleCloud
Platform/kubernetes/plugin/cmd/kube-scheduler]
W0717 16:05:26.742413 11344 server.go:83] Neither --kubeconfig nor --master was
specified. Using default API client. This might not work.
E0717 16:05:27.747413 11344 reflector.go:136] Failed to list *api.Node: Get
http://localhost:8080/api/v1/nodes?fieldSelector=spec.unschedulable%3Dfalse: dial
tcp 127.0.0.1:8080: ConnectEx tcp: No connection could be made because the target
machine actively refused it.
E0717 16:05:27.748413 11344 reflector.go:136] Failed to list *api.Pod: Get
http://localhost:8080/api/v1/pods?fieldSelector=spec.nodeName%21%3D: dial tcp
127.0.0.1:8080: ConnectEx tcp: No connection could be made because the target machine
actively refused it.
```

在 Kubernetes 的源码里包括不少单元测试，你可以在 LiteIDE 里运行通过，但有部分测试代码目前在 Windows 上无法通过，毕竟 Kubernetes 是为 Linux 打造的。接下来我们分析下 Kubernetes 源码的整体结构，Kubernetes 的源码总体分为 pkg、cmd、plugin、test 等顶级 package，其中 pkg 为 Kubernetes 的主体代码，cmd 为 Kubernetes 所有后台进程的代码（如 kube-apiserver 进程、kube-controller-manager 进程、kube-proxy 进程、kubelet 进程等），plugin 包含一些插件及 kuber-scheduler 的代码，test 为 Kubernetes 的一些测试代码。

从总体来看，Kubernetes 1.0 的当前包结构还是有点乱，开源团队还在继续优化中，可以从源码的 TODO 注释中看出这一点。表 6.1 给出了 Kubernetes 当前主要 package 的源码分析结果。

表 6.1 Kubernetes 当前主要 package 的源码分析结果

package	模 块 用 途	类 数 量
admission	权限控制框架，采用了责任链模式、插件机制	少
api	Kubernetes 所提供的 Rest API 接口的相关类，例如接口数据结构相关的 MetaData 结构、Volume 结构、Pod 结构、Service 结构等，以及数据格式验证转换工具类等，由于 API 是分版本的，所以这里是每个版本一个子 Package，例如 v1beta、v1 及 latest	中
apiserver	实现了 HTTP Rest 服务的一个基础性框架，用于 Kubernetes 的各种 Rest API 的实现，在 apiserver 包内也实现了 HTTP Proxy，用于转发请求（到其他组件，比如 Minion 节点上）	中
auth	3A 认证模块，包括用户认证、鉴权的相关组件	少

续表

package	模 块 用 途	类 数 量
client	是 Kubernetes 中公用的客户端部分的相关代码，实现协议为 HTTP Rest，用于提供一个具体的操作，例如对 Pod、Service 等的增删改查，这个模块也定义了 kubeletClient，同时为了高效地进行对象查询，此模块也实现了一个带缓存功能的存储接口 Store	多
cloudprovider	定义了云服务提供商运行 Kubernetes 所需的接口，包括 TCPLoadBalancer 的获取和创建；获取当前环境中的节点列表（节点是一个云主机）和节点的具体信息；获取 Zone 信息；获取和管理路由的接口等，默认实现了 AWS、GCE、Mesos、OpenStack、RackSpace 等云服务供应商的接口	中
controller	这部分提供了资源控制器的简单框架，用于处理资源的添加、变更、删除等事件的派发和执行，同时实现了 Kubernetes 的 ReplicationController 的具体逻辑	少
kubectrl	Kubernetes 的命令行工具 kubectrl 的代码模块，包括创建 Pod、服务、Pod 扩容、Pod 滚动升级等各种命令的具体实现代码	多
kubelet	Kubernetes 的 kubelet 的代码模块，是 Kubernetes 的核心模块之一，定义了 Pod 容器的接口，提供了 Docker 与 rkt 两种容器实现类，完成了容器及 Pod 的创建，以及容器状态的监控、销毁、垃圾回收等功能	多
master	Kubernetes 的 Master 节点代码模块，创建 NodeRegistry、PodRegistry、ServiceRegistry、EndpointRegistry 等组件，并且启动 Kubernetes 自身的相关服务，服务的 ClusterIP 地址分配及服务的 NodePort 端口分配，也是在这里完成的	少
proxy	Kubernetes 的服务代理和负载均衡相关功能的模块代码，目前实现了 round-robin 的负载均衡算法	少
registry	Kubernetes 的 NodeRegistry、PodRegistry、ReplicationControllerRegistry、ServiceRegistry、EndpointRegistry、PersistentVolumeRegistry 等注册表服务的接口及对应 Rest 服务的相关代码	多
runtime	为了让多个 API 版本共存，需要采用一些设计来完成不同 API 版本的数据结构的转换，API 中数据对象的 Encode/Decode 逻辑也最好集中化，Runtime 包就是为了这个目的而设计的	少
volume	实现了 Kubernetes 的各种 Volume 类型，分别对应亚马逊 ESB 存储、谷歌 GCE 的存储、Linux Host 目录存储、GlusterFS 存储、iSCSI 存储、NFS 存储、RBD 存储等，volume 包同时实现了 Kubernetes 容器的 Volume 卷的挂载、卸载功能	多
cmd	包括了 Kubernetes 所有后台进程的代码（如 kube-apiserver 进程、kube-controller-manager 进程、kube-proxy 进程、kubelet 进程等），而这些进程具体的业务逻辑代码则都在 pkg 中实现了	
plugin	子包 cmd/kuber-scheduler 实现了 Schedule Server 的框架，用于执行具体的 Scheduler 的调度，pkg/admission 子包则实现了 Admission 权限框架的一些默认实现类，例如 alwaysAdmit、alwaysDeny 等；pkg/auth 子包实现了权限认证框架（auth 包的）里定义的认证接口类，例如 HTTP BasicAuth、X509 证书认证；pkg/scheduler 子包则定义了一些具体的 Pod 调度器（Scheduler）	中

## 6.2 kube-apiserver 进程源码分析

Kubernetes API Server 是由 kube-apiserver 进程实现的，它运行在 Kubernetes 的管理节点——Master 上并对外提供 Kubernetes Restful API 服务，它提供的主要是与集群管理相关的 API 服务，例如校验 pod、service、replication controller 的配置并存储到后端的 etcd Server 上。下面我们分别对其启动过程、关键代码分析及设计总结等进行深入讲解。

### 6.2.1 进程启动过程

kube-apiserver 进程的入口类源码位置如下：

github.com/GoogleCloudPlatform/kubernetes/cmd/kube-apiserver/apiserver.go

入口 main() 函数的逻辑如下：

```
func main() {
    runtime.GOMAXPROCS(runtime.NumCPU())
    rand.Seed(time.Now().UTC().UnixNano())

    s := app.NewAPIServer()
    s.AddFlags(pflag.CommandLine)

    util.InitFlags()
    util.InitLogs()
    defer util.FlushLogs()

    verflag.PrintAndExitIfRequested()

    if err := s.Run(pflag.CommandLine.Args()); err != nil {
        fmt.Fprintf(os.Stderr, "%v\n", err)
        os.Exit(1)
    }
}
```

上述代码的核心为下面三行，创建一个 APIServer 结构体并将命令行启动参数传入，最后启动监听：

```
s := app.NewAPIServer()
s.AddFlags(pflag.CommandLine)
s.Run(pflag.CommandLine.Args())
```

我们先来看看都有哪些常用的命令行参数被传递给了 APIServer 对象，下面是运行在 Master 节点的 kube-apiserver 进程的命令行信息：

```
/usr/bin/kube-apiserver --logtostderr=true --etcd_servers=http://127.0.0.1:4001 --address=0.0.0.0 --port=8080 --kubelet_port=10250 --allow_privileged=false --service-cluster-ip-range=10.254.0.0/16
```

可以看到关键的几个参数有 `etcd_servers` 的地址、`APIServer` 绑定和监听的本地地址、`kubelet` 的运行端口及 `Kubernetes` 服务的 `clusterIP` 地址。

下面是 `app.NewAPIServer()` 的代码，我们看到这里的控制还是很全面的，包括安全控制（`CertDirectory`、`HTTPS` 默认启动）、权限控制（`AuthorizationMode`、`AdmissionControl`）、服务限流控制（`APIRate`、`APIBurst`）等，这些逻辑说明了 `APIServer` 是按照企业级平台的标准所设计和实现的。

```
func NewAPIServer() *APIServer {
    s := APIServer{
        InsecurePort:      8080,
        InsecureBindAddress: util.IP(net.ParseIP("127.0.0.1")),
        BindAddress:        util.IP(net.ParseIP("0.0.0.0")),
        SecurePort:         6443,
        APIRate:            10.0,
        APIBurst:           200,
        APISuffix:          "/api",
        EventTTL:           1 * time.Hour,
        AuthorizationMode:   "AlwaysAllow",
        AdmissionControl:    "AlwaysAdmit",
        EtcdPathPrefix:      master.DefaultEtcdPathPrefix,
        EnableLogsSupport:   true,
        MasterServiceNamespace: api.NamespaceDefault,
        ClusterName:         "kubernetes",
        CertDirectory:       "/var/run/kubernetes",

        RuntimeConfig: make(util.ConfigurationMap),
        KubeletConfig: client.KubeletConfig{
            Port:      ports.KubeletPort,
            EnableHttps: true,
            HTTPTimeout: time.Duration(5) * time.Second,
        },
    }

    return &s
}
```

创建了 `APIServer` 结构体实例后，`apiserver.go` 将此实例传入子包 `app/server.go` 的 `func(s *APIServer) Run(_ []string)` 方法里，最终绑定本地端口并创建一个 `HTTP Server` 与一个 `HTTPS Server`，从而完成整个进程的启动过程。

`Run` 方法的代码有很多，这里就不再列出源码，对该方法的源码解读如下。

- (1) 调用 `verifyClusterIPFlags` 方法，验证 `ClusterIP` 参数是否已设置及是否有效。
- (2) 验证 `etcd-servers` 的参数是否已设置。
- (3) 如果初始化 `CloudProvider`，且没有 `CloudProvider` 的参数，则日志告警并继续。
- (4) 根据 `KubeletConfig` 的配置参数，调用 `pkg/Client/kubeclient.go` 中的方法 `NewKubeletClient()` 创建一个 `kubelet Client` 对象，这其实是一个 `HTTPKubeletClient` 实例，目前只用于 `kubelet` 的健康检查（`KubeletHealthChecker`）。
- (5) 判断哪些 `API Version` 需要关闭，目前在 1.0 代码中默认关闭了 `v1beta3` 的 `API` 版本。
- (6) 创建一个 `Kubernetes` 的 `RestClient` 对象，具体的代码在 `pkg/client/helper.go` 的 `TransportFor()` 方法里完成，通过它完成 `Pod`、`Replication Controller` 及 `Kubernetes Service` 等对象的 `CRUD` 操作。
- (7) 创建用于访问 `etcd Server` 的客户端，具体代码在 `newEtcd()` 方法里实现，从代码调用中可以看出，`Kubernetes` 采用的是 `github.com/coreos/go-etcd/client.go` 这个客户端实现。
- (8) 建立鉴权（`Authenticator`）、授权（`Authorizer`）、服务许可框架和插件（`AdmissionControl`）的相关代码逻辑。
- (9) 获取和设置 `APIServer` 的 `ExternalHost` 的名称，如果没有提供 `ExternalHost` 参数，且 `Kubernetes` 运行在谷歌的 `GCE` 云平台上，则尝试通过 `CloudProvider` 接口获取本机节点的外部 `IP` 地址。
- (10) 如果运行在云平台中，则安装本机的 `SSH Key` 到 `Kubernetes` 集群中的所有虚拟机上。
- (11) 用 `APIServer` 的数据及上述过程中创建的一些对象（`KubeletClient`、`etcdClient`、`authenticator`、`admissionController` 等）作为参数，构造 `Kubernetes Master` 的 `Config` 结构（`pkg/master/master.go`），以此生成一个 `Master` 实例，具体代码在 `master.go` 中的 `New(c *Config)` 方法里。
- (12) 用上述创建的 `Master` 实例，分别创建 `HTTP Server` 及安全的 `HTTPS Server` 来开始监听客户端的请求，至此整个进程启动完毕。

## 6.2.2 关键代码分析

在 6.2.1 节里对 `kube-apiserver` 进程的启动过程进行了详细分析，我们发现 `Kubernetes API Service` 的关键代码就隐藏在 `pkg/master/master.go` 里，`APIServer` 这个结构体只不过是一个参数传递通道而已，它的数据最终传给了 `pkg/master/master.go` 里的 `Master` 结构体，下面是它的完整定义：

```
// Master contains state for a Kubernetes cluster master/api server.
type Master struct {
```



```

// "Inputs", Copied from Config
serviceClusterIPRange *net.IPNet
serviceNodePortRange  util.PortRange
cacheTimeout          time.Duration
minRequestTimeout     time.Duration

mux                    apiserver.Mux
muxHelper              *apiserver.MuxHelper
handlerContainer       *restful.Container
rootWebService         *restful.WebService
enableCoreControllers bool
enableLogsSupport      bool
enableUISupport        bool
enableSwaggerSupport   bool
enableProfiling        bool
apiPrefix              string
corsAllowedOriginList  util.StringList
authenticator          authenticator.Request
authorizer             authorizer.Authorizer
admissionControl       admission.Interface
masterCount            int
v1beta3                bool
v1                     bool
requestContextMapper   api.RequestContextMapper

// External host is the name that should be used in external (public internet)
URLs for this master
externalHost string
// clusterIP is the IP address of the master within the cluster.
clusterIP      net.IP
publicReadWritePort int
serviceReadWriteIP net.IP
serviceReadWritePort int
masterServices  *util.Runner

// storage contains the RESTful endpoints exposed by this master
storage map[string]rest.Storage
// registries are internal client APIs for accessing the storage layer
// TODO: define the internal typed interface in a way that clients can
// also be replaced
nodeRegistry      minion.Registry
namespaceRegistry namespace.Registry
serviceRegistry   service.Registry
endpointRegistry  endpoint.Registry
serviceClusterIPAllocator service.RangeRegistry

```

```
serviceNodePortAllocator service.RangeRegistry
// "Outputs"
    Handler      http.Handler
    InsecureHandler http.Handler

    // Used for secure proxy
    dialer      apiserver.ProxyDialerFunc
    tunnels     *util.SSHTunnelList
    tunnelsLock sync.Mutex
installSSHKey InstallSSHKey
    lastSync     int64 // Seconds since Epoch
    lastSyncMetric prometheus.GaugeFunc
    clock        util.Clock
}
```

在这段代码里，除了之前我们熟悉的那些变量，又多了几个陌生的重要变量，接下来我们逐一对其进行分析讲解。

首先是类型为 `apiserver.Mux`（来自文件 `pkg/apiserver/apiserver.go`）的 `mux` 变量，下面是对它的定义：

```
// mux is an object that can register http handlers.
type Mux interface {
    Handle(pattern string, handler http.Handler)
    HandleFunc(pattern string, handler func(http.ResponseWriter, *http.Request))
}
```

如果你熟悉 `Socket` 编程，特别使用过或者研究过 `HTTP Rest` 的一些框架，那么对于这个 `Mux` 接口就再熟悉不过了，它是一个 `HTTP` 的多分器（`Multiplexer`），其实它也是 `Golang HTTP` 基础包里的 `http.ServeMux` 的一个接口子集，用于派发（`Dispatch`）某个 `Request` 路径（这里用 `pattern` 变量表示）到对应的 `http.Handler` 进行处理。实际上在 `master.go` 代码中是生成一个 `http.ServeMux` 对象并赋值给 `apiserver.Mux` 变量，在代码中还有强制类型转换的语句。从上述分析来看，`apiserver.Mux` 的引入是设计的一个败笔，并没有增加什么价值，反而增加了理解代码的难度。此外，为了更好地实现 `Rest` 服务，`Kubernetes` 在这里引入了一个第三方的 `REST` 框架：[github.com/emicklei/go-restful](https://github.com/emicklei/go-restful)。

`go-restful` 在 `GitHub` 上有 36 个贡献者，采用了“路由”映射的设计思想，并且在 `API` 设计中使用了流行的 `Fluent Style` 风格，使用起来酣畅淋漓，也难怪 `Kubernetes` 选择了它。下面是 `go-restful` 的优良特性。

- ◎ `Ruby on Rails` 风格的 `Rest` 路由映射，例如 `/people/{person_id}/groups/{group_id}`。
- ◎ 大大简化了 `Rest API` 的开发工作。
- ◎ 底层实现采用 `Golang` 的 `HTTP` 栈，几乎没有限制。

- ◎ 拥有完整的单元包代码，很容易开发一个可测试的 Rest API。

- ◎ Google AppEngine ready。

go-restful 框架中的核心对象如下。

- ◎ restful.Container: 代表一个 HTTP Rest 服务器，包括一组 restful.WebService 对象和一个 http.ServeMux 对象，使用 RouteSelector 进行请求派发。
- ◎ restful.WebService: 表示一个 Rest 服务，由多个 Rest 路由 (restful.Route) 组成，这一组 Rest 路由共享同一个 Root Path。
- ◎ restful.Route: 表示一个 Rest 路由，Rest 路由主要由 Rest Path、HTTP Method、输入输出类型 (HTML/JSON) 及对应的回调函数 restful.RouteFunction 组成。
- ◎ restful.RouteFunction: 一个用于处理具体的 REST 调用的函数接口定义，具体定义为 type RouteFunction func(\*Request, \*Response)。

Master 结构体里包含了对 restful.Container 与 restful.WebService 这两个 go-restful 核心对象的引用，在接下来的 Master 对象的构造方法中（对应代码为 master.go 的 func New(c \*Config) \*Master）被初始化。那么，问题又来了，Kubernetes 的这么一堆 Rest API 又是在哪里定义的，是如何被绑定到 restful.Route 里的呢？

要理解这个问题，我们要首先弄清楚 Master 结构体中的变量：

```
storage map[string]rest.Storage
```

storage 变量是一个 Map，Key 为 Rest API 的 path，Value 为 rest.Storage 接口，此接口是一个通用的符合 Restful 要求的资源存储服务接口，每个服务接口负责处理一类 (Kind) Kubernetes Rest API 中的数据对象——资源数据，只有一个接口方法：New()，New() 方法返回该 Storage 服务所能识别和管理的某种具体的资源数据的一个空实例。

```
type Storage interface {
    New() runtime.Object
}
```

在运行期间，Kubernetes API Runtime 运行时框架会把 New() 方法返回的空对象的指针传入 Codec.DecodeInto([]byte, runtime.Object) 方法中，从而完成 HTTP Rest 请求中的 Byte 数组反序列化逻辑。Kubernetes API Server 中所有对外提供服务的 Restful 资源都实现了此接口，这些资源包括 pods、bindings、podTemplates、replicationControllers、services 等，完整的列表就在 master.go 的 func (m \*Master) init(c \*Config) 中，下面是相关代码片段（截取部分代码）。

```
m.storage = map[string]rest.Storage{
    "pods":          podStorage.Pod,
    "pods/status":   podStorage.Status,
    "pods/log":      podStorage.Log,
```

```

    "pods/exec":      podStorage.Exec,
    "pods/portforward": podStorage.PortForward,
    "pods/proxy":     podStorage.Proxy,
    "pods/binding":   podStorage.Binding,
    "bindings":       podStorage.Binding,

    "podTemplates": podTemplateStorage,

    "replicationControllers": controllerStorage,
    "services":               service.NewStorage(m.serviceRegistry,
m.nodeRegistry, m.endpointRegistry, serviceClusterIPAllocator, serviceNodePort
Allocator, c.ClusterName),
    "endpoints":           endpointsStorage,
    "minions":             nodeStorage,

```

看到上面这段代码，你在潜意识里已经明白，这其实就是似曾相识的 **Kubernetes Rest API** 列表，**storage** 这个 Map 的 Key 就是 Rest API 的访问路径，Value 却不是之前说好的 **restful.Route**。聪明的你一定想到了答案：必然存在一个“转换适配”的方法来实现上述转换！这就是 **pkg/apiserver/api\_installer.go** 包里那个逻辑清晰但源码超长的方法 **registerResourceHandlers** 了：

```

func (a *APIInstaller) registerResourceHandlers(path string, storage rest.
Storage, ws *restful.WebService, proxyHandler http.Handler)

```

上述方法把一个 **path** 对应的 **rest.Storage** 转换成一系列的 **restful.Route** 并添加到指针 **restful.WebService** 中。这个函数的代码之所以很长，是因为有各种情况要考虑，比如 **pods/portforward** 这种路径要处理 **child**，还要判断每种 **Storage** 资源类型所支持的操作类型；比如是否支持 **create**、**delete**、**update** 及是否支持 **list**、**watch**、**patcher** 操作等，对各种情况都考虑以后，这个函数的代码量已接近 500 行！估计 **Kubernetes** 这段代码的作者也不大好意思，于是外面封装了简单函数：**func(a \*APIInstaller)Install**，内部循环调用 **registerResourceHandlers**，返回最终的 **restful.WebService** 对象，此方法的主要代码如下：

```

// Installs handlers for API resources.
func (a *APIInstaller) Install() (ws *restful.WebService, errors []error) {
    // Register the paths in a deterministic (sorted) order to get a deterministic
swagger spec.
    paths := make([]string, len(a.group.Storage))
    var i int = 0
    for path := range a.group.Storage {
        paths[i] = path
        i++
    }
    sort.Strings(paths)
    for _, path := range paths {
        if err := a.registerResourceHandlers(path, a.group.Storage[path], ws,
proxyHandler); err != nil {

```

```

        errors = append(errors, err)
    }
}
return ws, errors
}

```

为了区分 API 的版本，在 `apiserver.go` 里定义了一个结构体：APIGroupVersion。以下是其代码：

```

type APIGroupVersion struct {
    Storage map[string]rest.Storage
    Root    string
    Version string
    // ServerVersion controls the Kubernetes APIVersion used for common objects
in the apiserver
    // schema like api.Status, api.DeleteOptions, and api.ListOptions. Other
implementors may
    // define a version "v1beta1" but want to use the Kubernetes "v1beta3" internal
objects. If
    // empty, defaults to Version.
    ServerVersion string

    Mapper meta.RESTMapper

    Codec      runtime.Codec
    Typer      runtime.ObjectTyper
    Creator    runtime.ObjectCreator
    Convertor  runtime.ObjectConvertor
    Linker     runtime.SelfLinker

    Admit      admission.Interface
    Context    api.RequestContextMapper

    ProxyDialerFn ProxyDialerFunc
    MinRequestTimeout time.Duration
}

```

我们注意到 APIGroupVersion 是与 rest.Storage Map 捆绑的，并且绑定了相应版本的 Codec、Convertor 用于版本转换，这样就很容易理解 Kubernetes 是怎样区分多版本 API 的 Rest 服务的。以下是过程详解。

首先，在 APIGroupVersion 的 InstallREST(container \*restful.Container)方法里，用 Version 变量来构造一个 Rest API Path 前缀并赋值给 APIInstaller 的 prefix 变量，并调用它的 Install()方法完成 Rest API 的转换，代码如下：

```

func (g *APIGroupVersion) InstallREST(container *restful.Container) error {
    info := &APIRequestInfoResolver{util.NewStringSet(strings.TrimPrefix(g.Root,
"/"), g.Mapper)
}

```

```
prefix := path.Join(g.Root, g.Version)
installer := &APIInstaller{
    group:      g,
    info:       info,
    prefix:     prefix,
    minRequestTimeout: g.MinRequestTimeout,
    proxyDialerFn:  g.ProxyDialerFn,
}
ws, registrationErrors := installer.Install()
container.Add(ws)
```

接着，在 `APIInstaller` 的 `Install()`方法里用 `prefix`（API 版本）前缀生成 `WebService` 的相对根路径：

```
func (a *APIInstaller) newWebService() *restful.WebService {
    ws := new(restful.WebService)
    ws.Path(a.prefix)
    ws.Doc("API at"+ a.prefix +"version"+ a.group.Version)
    // TODO: change to restful.MIME_JSON when we set content type in client
    ws.Consumes("*/")
    ws.Produces(restful.MIME_JSON)
    ws.ApiVersion(a.group.Version)

    return ws
}
```

最后，在 `Kubernetes` 的 `Master` 初始化方法 `func (m *Master) init (c *Config)`里生成不同的 `APIGroupVersion` 对象，并调用 `InstallRest()`方法，完成最终的多版本 API 的 Rest 服务装配流程：

```
if m.v1beta3 {
    if err := m.api_v1beta3().InstallREST(m.handlerContainer); err != nil {
        glog.Fatalf("Unable to setup API v1beta3: %v", err)
    }
    apiVersions = append(apiVersions, "v1beta3")
}
if m.v1 {
    if err := m.api_v1().InstallREST(m.handlerContainer); err != nil {
        glog.Fatalf("Unable to setup API v1: %v", err)
    }
    apiVersions = append(apiVersions, "v1")
}
```

至此，Rest API 的多版本问题还有最后一个需要澄清，即在不同的版本中接口的输入输出参数的格式是有差别的，`Kubernetes` 是怎么处理这个问题的？

要弄明白这一点，我们首先要研究 Kubernetes API 里的数据对象的序列化、反序列化的实现机制。为了同时解决数据对象的序列化、反序列化与多版本数据对象的兼容和转换问题，Kubernetes 设计了一套复杂的机制，首先，它设计了 `conversion.Scheme` 这个结构体（`pkg/conversion/schema.go` 里），以下是对它的定义：

```
// Scheme defines an entire encoding and decoding scheme.
type Scheme struct {
    // versionMap allows one to figure out the go type of an object          //with
    the given version and name.
    versionMap map[string]map[string]reflect.Type
    // typeToVersion allows one to figure out the version for a given //go object
    The reflect.Type we index by should *not* be a pointer. If the same type
    // is registered for multiple versions, the last one wins.
    typeToVersion map[reflect.Type]string
    // typeToKind allows one to figure out the desired "kind" field //for a given
    go object. Requirements and caveats are the same as typeToVersion.
    typeToKind map[reflect.Type][]string
    // converter stores all registered conversion functions. It also //has default
    covertng behavior.
    converter *Converter
    // cloner stores all registered copy functions. It also has default
    // deep copy behavior.
    cloner *Cloner
    // Indent will cause the JSON output from Encode to be indented, iff it is true.
    Indent bool
    // InternalVersion is the default internal version. It is recommended that
    // you use "" for the internal version.
    InternalVersion string
    // MetaInsertionFactory is used to create an object to store and retrieve
    // the version and kind information for all objects. The default // uses
    the keys "apiVersion" and "kind" respectively.
    MetaFactory MetaFactory
}
```

在上述代码中可以看到，`typeToVersion` 与 `versionMap` 属性是为了解决数据对象的序列化与反序列化问题，`converter` 属性则负责不同版本的数据对象转换问题，Kubernetes 这个设计思路简单、方便地解决了多版本的序列化和数据转换问题，让人不得不赞！下面是 `conversion.Scheme` 里序列化、反序列化的核心方法 `NewObject()` 的代码：通过查找 `versionMap` 里匹配的注册类型，以反射方式生成一个空的数据对象：

```
func (s *Scheme) NewObject(versionName, kind string) (interface{}, error) {
    if types, ok := s.versionMap[versionName]; ok {
        if t, ok := types[kind]; ok {
            return reflect.New(t).Interface(), nil
        }
    }
}
```

```

    return nil, &notRegisteredErr{kind: kind, version: versionName}
}
return nil, &notRegisteredErr{kind: kind, version: versionName}
}

```

而 `pkg/conversion/encode.go` 与 `decode.go` 则在 `conversion.Scheme` 提供的基础功能之上，完成了最终的序列化、反序列化功能。下面是 `encode.go` 里的主方法 `EncodeToVersion(..)` 的关键代码片段：

```

//确定要转换的源对象的版本号 and 类别
objVersion, objKind, err := s.ObjectVersionAndKind(obj) 象
//生成目标版本的空对象
objOut, err := s.NewObject(destVersion, objKind)
//生成转换过程中所需的 Metadata 信息
flags, meta := s.generateConvertMeta(objVersion, destVersion, obj)
//调用 converter 的方法将源对象的数据填充到目标对象 objOut
err = s.converter.Convert(obj, objOut, flags, meta)
//用 JSON 将目标对象转换成 byte[] 数组，完成序列化过程
data, err = json.Marshal(obj)

```

更进一步，Kubernetes 在 `conversion.Scheme` 的基础上又做了一个封装工具类 `runtime.Scheme`，可以看作前者的代理类，主要增加了 `fieldLabelConversionFuncs` 这个 Map 属性，用于解决数据对象的属性名称的兼容性转换和校验，比如将需要兼容 Pod 的 `spec.host` 属性改为 `spec.nodeName` 的情况。

注意到 `conversion.Scheme` 只是实现了一个序列化与类型转换的框架 API，提供了注册资源数据类型与转换函数的功能，那么具体的资源数据对象类型、转换函数又是在哪个包里实现的呢？答案是 `pkg/api`。Kubernetes 为不同的 API 版本提供了独立的数据类型和相关的转换函数并按照版本号命名 Package，如 `pkg/api/v1`、`pkg/api/v1beta3` 等，而当前默认版本（内部版本）则存在于 `pkg/api` 目录下。

以 `pkg/api/v1` 为例，在每个目录里都包括如下关键源码：

- ◎ `types.go` 定义了 Rest API 接口里所涉及的所有数据类型，v1 版本有 2000 行代码；
- ◎ 在 `conversion.go` 与 `conversion_generated.go` 里定义了 `conversion.Scheme` 所需的从内部版本到 v1 版本的类型转换函数，其中 `conversion_generated.go` 中的代码有 5000 行之多，当然这是通过工具自动生成的代码；
- ◎ `register.go` 负责将 `types.go` 里定义的数据类型与 `conversion.go` 里定义的数据转换函数注册到 `runtime.Schema` 里。

`pkg/api` 里的 `register.go` 初始化生成并持有一个全局的 `runtime.Scheme` 对象，并将当前默认版本的数据类型（`pkg/api/types.go`）注册进去，相关代码如下：

```
var Scheme = runtime.NewScheme()
```



```
func init() {
    Scheme.AddKnownTypes("",
        &Pod{},
        &PodList{},
        &PodStatusResult{},
        &PodTemplate{},
        &PodTemplateList{},
        &ReplicationControllerList{},
//此次省略 30 多个数据类型
        &ServiceList{},
        &Service{},
        &NodeList{},
        &Node{},
//省略

```

而 `pkg/api/v1/register.go` 与 `v1beta3` 下的 `register.go` 在初始化过程中分别把与版本相关的数据类型和转换函数注册到全局的 `runtime.Scheme` 中：

```
func init() {
    // Check if v1 is in the list of supported API versions.
    if !registered.IsRegisteredAPIVersion("v1") {
        return
    }

    // Register the API.
    addKnownTypes()
    addConversionFuncs()
    addDefaultingFuncs()
}

```

这样一来，其他地方都可以通过 `runtime.Scheme` 这个全局变量来完成 Kubernetes API 中的数据对象的序列化和反序列化逻辑了，比如 Kubernetes API Client 包就大量使用了它，下面是 `pkg/client/pods.go` 里 Pod 删除的 `Delete()` 方法的代码：

```
// Delete takes the name of the pod, and returns an error if one occurs
func (c *Pods) Delete(name string, options *api.DeleteOptions) error {
    // TODO: to make this reusable in other client libraries
    if options == nil {
        return c.r.Delete().Namespace(c.ns).Resource("pods").Name(name).
Do().Error()
    }
    body, err := api.Scheme.EncodeToVersion(options, c.r.APIVersion())
    if err != nil {
        return err
    }
    return c.r.Delete().Namespace(c.ns).Resource("pods").Name(name).
Body(body).Do().Error()
}

```

清楚了 Kubernetes Rest API 中的数据对象的序列化机制及多版本的实现原理之后，我们接着分析下面这个重要流程的实现细节。

Kubernetes 中实现了 `rest.Storage` 接口的服务在转换成 `restful.RouteFunction` 以后，是怎样处理一个 Rest 请求并最终完成基于后端存储服务 etcd 上的具体操作过程的？

首先，Kubernetes 设计了一个名为“注册表”的 Package (`pkg/registry`)，这个 Package 按照 `rest.Storage` 服务所管理的资源数据的类型而划分为不同的子包，每个子包都由相同命名的一组 Golang 代码来完成具体的 Rest 接口的实现逻辑。

下面我们以 Pod 的 Rest 服务实现为例，其与“注册表”相关的代码位于 `pkg/registry/pod` 中，在 `registry.go` 里定义了 Pod 注册表服务的接口：

```
type Registry interface {
    // ListPods obtains a list of pods having labels which match selector.
    ListPods(ctx api.Context, label labels.Selector) (*api.PodList, error)
    // Watch for new/changed/deleted pods
    WatchPods(ctx api.Context, label labels.Selector, field fields.Selector,
resourceVersion string) (watch.Interface, error)
    // Get a specific pod
    GetPod(ctx api.Context, podID string) (*api.Pod, error)
    // Create a pod based on a specification.
    CreatePod(ctx api.Context, pod *api.Pod) error
    // Update an existing pod
    UpdatePod(ctx api.Context, pod *api.Pod) error
    // Delete an existing pod
    DeletePod(ctx api.Context, podID string) error
}
```

我们看到这个 Pod 注册表服务是针对 Pod 的 CRUD 的操作接口的一个定义，在入口参数中除了调用的上下文环境 `api.Context`，就是我们之前分析过的 `pkg/api` 包中的 `Pod` 这个资源数据对象。为了实现强类型的方法调用，在 `registry.go` 里定义了一个名为 `storage` 的结构体，`storage` 实现 `Registry` 接口，可以看作一种代理设计模式，因为具体的操作都是通过内部 `rest.StandardStorage` 来实现的。下面是截取的 `registry.go` 中的 `create`、`update`、`delete` 的源码：

```
func (s *storage) CreatePod(ctx api.Context, pod *api.Pod) error {
    _, err := s.Create(ctx, pod)
    return err
}

func (s *storage) UpdatePod(ctx api.Context, pod *api.Pod) error {
    _, _, err := s.Update(ctx, pod)
    return err
}
```

```
func (s *storage) DeletePod(ctx api.Context, podID string) error {
    _, err := s.Delete(ctx, podID, nil)
    return err
}
```

那么,这个实现了 `rest.StandardStorage` 通用接口的真正 `Storage` 又是什么?从 `Master` 对象的初始化函数中,我们发现了下面的相关代码:

```
func (m *Master) init(c *Config) {
    healthzChecks := []healthz.HealthzChecker{}
    m.clock = util.RealClock{}
    podStorage := podetcd.NewStorage(c.EtcdHelper, c.KubeletClient)
    podRegistry := pod.NewRegistry(podStorage.Pod)
```

`Master` 对象创建了一个私有变量 `podStorage`,其类型为 `PodStorage` (`pkg/registry/pod/etcd/etcd.go`),`Pod` 注册表服务实例 (`podRegistry`) 里真正的 `Storage` 是 `podStorage.Pod`。下面是 `podetcd` 的函数 `NewStorage` 中的关键代码:

```
func NewStorage(h tools.EtcdHelper, k client.ConnectionInfoGetter) PodStorage {
    store := &etcdgeneric.Etcd{
        NewFunc:      func() runtime.Object { return &api.Pod{} },
        NewListFunc:   func() runtime.Object { return &api.PodList{} },
        .....
    }
    return PodStorage{
        Pod:           &REST{*store},
        Binding:        &BindingREST{store: store},
        Status:         &StatusREST{store: &statusStore},
        Log:            &LogREST{store: store, kubeletConn: k},
        Proxy:          &ProxyREST{store: store},
        Exec:           &ExecREST{store: store, kubeletConn: k},
        PortForward:    &PortForwardREST{store: store, kubeletConn: k},
    }
}
```

在上述代码中我们看到:位于 `pkg/registry/generic/etcd/etcd.go` 里的 `etcd` 才是真正的 `Storage` 实现。而具体操作 `etcd` 的代码是靠 `tools.EtcdHelper` 这个类完成的,通过分析 `etcd.go` 里的 `func (e *Etcd)Create(ctx api.Context, obj runtime.Object)` 方法,我们知道创建一个 `etcd` 里的键值对的关键逻辑如下。

- ◎ 获取对象的名字: `name, err := e.ObjectNameFunc(obj)`。
- ◎ 获取 Key: `key, err := e.KeyFunc(ctx, name)`。
- ◎ 生成一个空的 `Object` 对象: `out := e.NewFunc()`。
- ◎ 将键值对写入 `etcd`: 在 `e.Helper.CreateObj(key, obj, out, ttl)` 方法中通过调用 `runtime.Codec` 完成从对象到字符串的转换,最终保存到 `etcd` 中。
- ◎ 回调创建完成后的处理逻辑: `e.AfterCreate(out)`。

注意到之前 PodStorage 创建 store 时重载了 ObjectNameFunc()、KeyFunc()、NewFunc()等函数，于是完成了针对 Pod 的创建过程，Kubernetes API 服务中的其他数据对象也都遵循同样的设计模式。

进一步研究代码，我们发现 PodStorage 中的 Pod、Binding、Status 等属性是 pkg/api/rest/rest.go 中几个不同的 Rest 接口的实现，并且通过 etcdgeneric.Etcd 这个实例来完成 Pod 的一些具体操作，比如这里的 StatusREST。下面是其相关代码片段：

```
// StatusREST implements the REST endpoint for changing the status of a pod.
type StatusREST struct {
    store *etcdgeneric.Etcd
}
// New creates a new pod resource
func (r *StatusREST) New() runtime.Object {
    return &api.Pod{}
}
// Update alters the status subset of an object.
func (r *StatusREST) Update(ctx api.Context, obj runtime.Object) (runtime.Object,
bool, error) {
    return r.store.Update(ctx, obj)
}
```

表 6.2 展现了 PodStorage 中的各个 XXXREST 接口与 pkg/api/rest/rest.go 里的相关 Rest 接口的一一对应关系。

表 6.2 PodStorage 中的各个 XXXREST 接口与 pkg/api/rest/rest.go 里的相关 Rest 接口的一一对应关系

PodStorage Rest 接口	对应 API Rest 框架的接口	接 口 功 能
REST	rest.Redirector rest.CreaterUpdater rest.Lister rest.Watcher rest.GracefulDeleter rest.Getter	重定向资源的路径 资源创建、更新接口 资源列表查询接口 Watcher 资源变化接口 支持延迟的资源删除接口 获取具体资源的信息接口
BindingREST	rest.Creater	创建资源的接口
StatusREST	Rest.Updater	更新资源的接口
LogREST	rest.GetterWithOptions	获取资源的接口
ExecREST\ProxyREST\ PortForwardREST	rest.Connector	连接资源的接口

其中 PodStorage.REST 接口究竟实现了哪些 API Rest 接口，这个比较隐晦，笔者也花费了一些时间来研究这个问题，这涉及 Go 语言的一个特殊特性：结构体内嵌一个其他类型的结构体指针，就可以使用内嵌结构体的方法，相当于面向对象语言中的“继承”。而 PodStorage.REST 恰恰嵌套了 etcdgeneric.Etcd 类型的匿名指针：&REST{\*store}，而 etcdgeneric.Etcd 则实现了

rest.Creator、rest.Lister、rest.Watcher 等资源管理接口的所有方法，PodStorage.REST 也“继承”了这些接口。

我们回头看看下面这段来自 api\_installer.go 的 registerResourceHandlers 函数中的片段：

```

creator, isCreator := storage.(rest.Creator)
namedCreator, isNamedCreator := storage.(rest.NamedCreator)
lister, isLister := storage.(rest.Lister)
getter, isGetter := storage.(rest.Getter)
getterWithOptions, isGetterWithOptions := storage.(rest.GetterWithOptions)
deleter, isDeleter := storage.(rest.Deleter)
gracefulDeleter, isGracefulDeleter := storage.(rest.GracefulDeleter)
updater, isUpdater := storage.(rest.Updater)
patcher, isPatcher := storage.(rest.Patcher)
watcher, isWatcher := storage.(rest.Watcher)
_, isRedirector := storage.(rest.Redirector)
connector, isConnector := storage.(rest.Connector)
storageMeta, isMetadata := storage.(rest.StorageMetadata)

```

上述代码对 storage 对象进行判断，以确定并标记它所满足的 API Rest 接口类型，而接下来的这段代码在此基础上确定此接口所包含的 actions，后者则对应到某种 HTTP 请求方法（GET/POST/PUT/DELETE）或者 HTTP PROXY、WATCH、CONNECT 等动作：

```

actions = appendIf(actions, action{"GET", itemPath, nameParams, namer}, isGetter)
actions = appendIf(actions, action{"PATCH", itemPath, nameParams, namer},
isPatcher)
actions = appendIf(actions, action{"DELETE", itemPath, nameParams, namer},
isDeleter)
actions = appendIf(actions, action{"WATCH", "watch/" + itemPath, nameParams,
namer}, isWatcher)
actions = appendIf(actions, action{"PROXY", "proxy/" + itemPath + "{path:*}",
proxyParams, namer}, isRedirector)
actions = appendIf(actions, action{"CONNECT", itemPath, nameParams, namer},
isConnector)

```

我们注意到 rest.Redirector 类型的 storage 被当作 PROXY 进行处理，由 apiserver.ProxyHandler 进行拦截，并调用 rest.Redirector 的 ResourceLocation 方法获取到资源的处理路径（可能包括一个非空的 http.RoundTripper，用于处理执行 Redirector 返回的 URL 请求）。Kubernetes API Server 中 PROXY 请求存在的意义在于透明地访问其他某个节点（比如某个 Minion）上的 API。

最后，我们来分析 registerResourceHandlers 中完成从 rest.Storage 到 restful.Route 映射的最后一段关键代码。下面是 rest.Getter 接口的 Storage 的映射代码：

```

case "GET": // Get a resource.
var handler restful.RouteFunction
handler = GetResource(getter, reqScope)

```

```
doc := "read the specified " + kind
route := ws.GET(action.Path).To(handler).Filter(m).Doc(doc).
Param(ws.QueryParameter("pretty", "If 'true', then the output is pretty printed.
")).
Operation("read"+namespaced+kind+strings.Title(subresource)).
Produces(append(storageMeta.ProducesMIMETypes(action.Verb), "application/
json")...).
Returns(http.StatusOK, "OK", versionedObject).Writes(versionedObject)

addParams(route, action.Params)
ws.Route(route)
```

上述代码首先通过函数 `GetResource()` 创建了一个 `restful.RouteFunction`，然后生成一个 `restful.route` 对象，最后注册到 `restful.WebService` 中，从而完成了 `rest.Storage` 到 Rest 服务的“最后一公里”通车。`GetResource()` 函数存在于 `pkg/apiserver/resthandler.go` 里，`resthandler.go` 提供了各种具体的 `restful.RouteFunction` 的实现函数，是真正触发 `rest.Storage` 调用的地方。下面是 `GetResource()` 方法的主要代码，可以看出这里是调用 `rest.Getter` 接口的 `Get()` 方法以返回某个资源对象：

```
func GetResource(r rest.Getter, scope RequestScope) restful.RouteFunction {
    return getResourceHandler(scope,
        func(ctx api.Context, name string, req *restful.Request) (runtime.Object,
error) {
            return r.Get(ctx, name)
        })
}
```

看了上面的代码，你可能会会有一个疑问：“说好的权限控制呢？”别急，看看下面的资源创建的 `createHandler()` 代码：

```
if admit.Handles(admission.Create) {
    userInfo, _ := api.UserFrom(ctx)
    err = admit.Admit(admission.NewAttributesRecord(obj, scope.Kind,
namespace, name, scope.Resource, scope.Subresource, admission.Create, userInfo))
    if err != nil {
        errorJSON(err, scope.Codec, w)
        return
    }
}
```

资源的 `Update`、`Delete`、`Connect`、`Patch` 等操作都有类似的权限控制，从 `Admit` 的参数 `admission.Attributes` 的属性来看，第三方系统可以开发细粒度的权限控制插件，针对任意资源的任意属性进行细粒度的权限控制，因为资源对象本身都传递到参数中了。

对 Kubernetes Rest API Server 的复杂实现机制和调用流程的总结如下。

- ◎ 在 `pkg/api` 包里定义了 Rest API 中涉及的资源对象、提供的 Rest 接口、类型转换框架和具体转换函数、序列化反序列化等代码。其中，资源对象和转换函数按照版本分包，

形成了 Kubernetes API Server 基础的框架，其中核心是各类资源（如 Node、Pod、PodTemplate、Service 等）及这些资源对应的 rest.Storage（Rest API 接口）。

- ◎ 在 pkg/runtime 包里最重要的对象是 Schema，它保存了 Kubernetes API Service 中注册的资源对象类型、转换函数等重要基础数据。另外，runtime 包也提供了获取 json/yaml 序列化、反序列化的 Codec 结构体，runtime 总体上与 pkg/api 密切相关，分离出来的目的是供其他模块方便使用。
- ◎ pkg/registry 包其实是把 pkg/api 中定义的各种资源对象所提供的 Rest 接口进一步规范定义并且实现对应的接口，其中 generate/etcd/etcd.go 里的 etcd 对象是一个真正实现了 rest.Storage 接口的基于 etcd 后端存储的服务框架，并且 Kubernetes 中的各种资源对象的具体 Storage 实现也是通过它来完成真正的“后端存储操作”。
- ◎ Kubernetes 采用了 go-restful 这个第三方的 Rest 框架，大大简化了 Rest 服务的开发，主要代码在 pkg/apiserver 源码包里。通过 APIGroupVersion 这个结构体可完成不同 API 版本的 Rest 路径映射，而 api\_installer.go 则实现了从 Kubernetes Rest.Storage 接口到 go-restful 的映射连接逻辑，对应 rest.Storage 的具体 restful.RouteFunction 则在 resthandler.go 里实现。

### 6.2.3 设计总结

如果你耐心看完了上面的每一段文字和代码，而且尝试追踪源码来加深对 6.2.1 节内容的理解，那么笔者相信你对于 Kubernetes API Server 的设计的第一个评价就是：“太复杂、太反常了！不就是一个 Rest Server 么，如果用 Java 语言，我可以分分钟搞定一个！”当然，你肯定有以下或者更多的假设。

- ◎ 放弃多版本 API 的兼容需求。
- ◎ 只采用一个特定的后端存储实现。
- ◎ API 只接收一种输入输出格式，比如 JSON 或者 YAML，而不是两种或更多。
- ◎ 放弃 Watch 这种高难度的 API。
- ◎ 不实现 Proxy 代理。
- ◎ 不做可拔插的权限控制设计（或者根本没有）。
- ◎ 每新增一种资源类型，就从头写很多代码来实现该资源的 Rest 服务。

虽然代码很复杂，但我们不得不承认，Kubernetes API Server 是一个精心“设计”的系统。

什么样的设计是一个好的设计？这个问题没有标准答案，但有一点是大家都认可的：好的设计要尽量提供一种好的框架机制，方便未来增加新功能或者自定义扩展某些特性。我们以这





考虑到大多数资源对象都需要基本的 CRUD 接口，这就是 `rest.StandardStorage` 这个聚合型“标准存储服务”接口出现的原因。而作为 `StandardStorage` 的默认实现，`pkg/registry/generic/etcd/etcd.go` 里 `etcd` 这个对象实现了基于 `etcd` 后端存储的所有具体操作，而各种资源的 `Storage` 服务则通过将请求代理到 `etcd` 对象上来完成具体的功能。

这里有点让人难以理解的是 `PodStorage` 与它的属性 `Pod` 的关系，其实 `PodStorage` 这个对象是一个聚合了与 `Pod` 相关的各个资源的存储服务，多看一下它的定义就能立刻明白了：

```
// PodStorage includes storage for pods and all sub resources
type PodStorage struct {
    Pod          *REST
    Binding      *BindingREST
    Status       *StatusREST
    Log          *LogREST
    Proxy        *ProxyREST
    Exec         *ExecREST
    PortForward  *PortForwardREST
}
```

所以，这里的 `PodStorage` 应该重命名为 `AllPodResStorage`，而真正的 `PodStorage` 上就是里面的那个 `Pod` 变量，这个变量是对 `etcd` 实例的一个引用，然后又实现了 `rest.Redirector` 接口。现在你终于能理解 `PodRegistry` 引用 `Pod` 变量而不是 `PodStorage` 来实现 `Pod` 操作的真正原因了吧？

最后，我们来说说 `PodRegistry` 存在的目的。从之前的代码分析来看，一个来自外部的针对某个资源的 `Rest API` 发起的请求最后落到对应资源的 `rest.Storage` 对象上，由 `restful.RouteFunction` 调用此对象的相关方法完成资源的操作并生成应答返回给客户端，这个过程并没有涉及对应资源的 `Registry` 服务。那么问题来了，资源的 `Registry` 接口存在的理由是什么呢？答案很简单，对比 `Storage` 接口与 `Registry` 中的资源创建方法的签名，下面是二者的源码对比，后者更符合“手工调用”：

```
Storage 中创建通用的资源对象的接口
Create(ctx api.Context, obj runtime.Object) (runtime.Object, error)
PodRegistry 中创建 Pod 资源的接口
CreatePod(ctx api.Context, pod *api.Pod) error
```

在 `Kubernetes API Server` 中为每类资源都创建并提供了一个 `Registry` 接口服务的目的是供内部模块的编程使用，而非对外提供服务，很多文档都错误理解了这个问题。

本节最后给出了如图 6.3 所示的经典的 `Kubernetes` 的 `Master` 节点数据流图，此刻这个图在你眼里可能已经什么都不算了，因为你已经洞穿了幕后的一切。

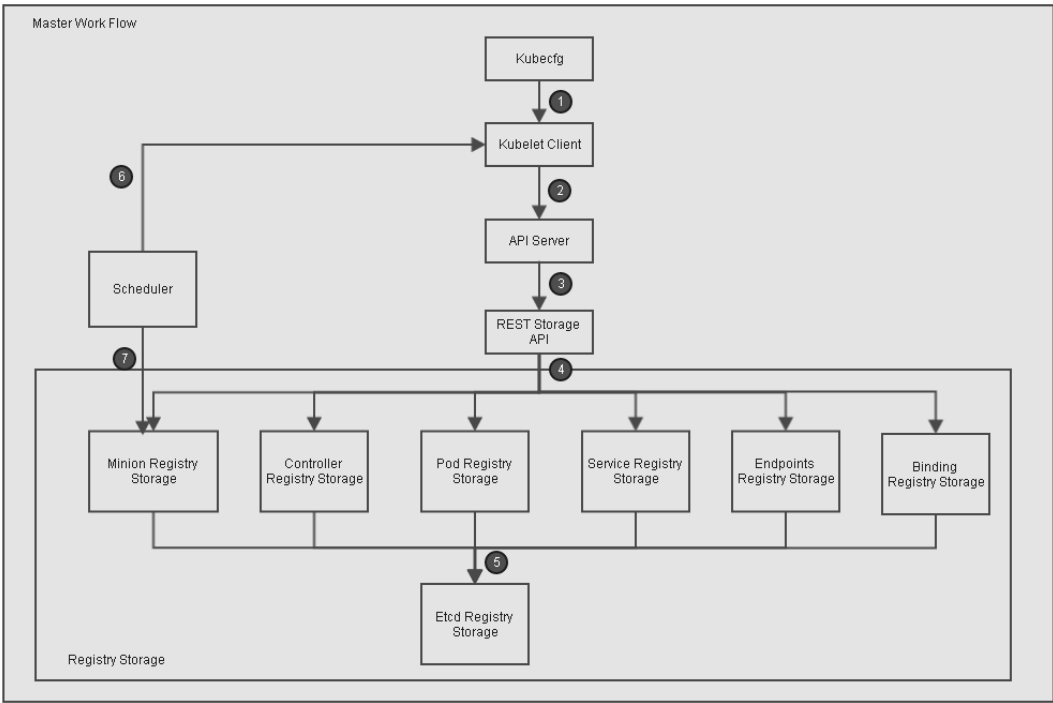


图 6.3 Master 节点数据流图

## 6.3 kube-controller-manager 进程源码分析

运行在 Master 节点上的第 2 个进程就是 kube-controller-manager 进程，即 Controller Manager Server，Kubernetes 的核心进程之一，其主要目的是实现 Kubernetes 集群的故障检测和恢复的自动化工作，比如内部组件 EndpointController 控制器负责 Endpoints 对象的创建和更新；ReplicationManager 根据注册表中的 ReplicationController 的定义，完成 Pod 的复制或者移除，以确保复制数量的一致性；NodeController 负责 Minion 节点的发现、管理和监控。

### 6.3.1 进程启动过程

kube-controller-manager 进程的入口源码位置如下：

```
github.com/GoogleCloudPlatform/kubernetes/cmd/kube-controller-manager/controller-manager.go
```

入口 `main()` 函数的逻辑如下：

```
func main() {
    runtime.GOMAXPROCS(runtime.NumCPU())
    s := app.NewCMServer()
    s.AddFlags(pflag.CommandLine)
    util.InitFlags()
    util.InitLogs()
    defer util.FlushLogs()
    verflag.PrintAndExitIfRequested()
    if err := s.Run(pflag.CommandLine.Args()); err != nil {
        fmt.Fprintf(os.Stderr, "%v\n", err)
        os.Exit(1)
    }
}
```

从源码可以看出，关键代码只有两行，创建一个 `CMServer` 并调用 `Run` 方法启动服务。下面我们分析 `CMServer` 这个结构体，它是 `Controller Manager Server` 进程的主要上下文数据结构，存放一些关键参数，表 6.3 是对 `CMServer` 中关键参数的解释。

表 6.3 对 `CMServer` 中关键参数的解释

属 性 名	默 认	含 义
<code>ConcurrentEndpointSyncs</code>	5s	并发执行的 Endpoint 的同步任务的数量
<code>ConcurrentRCSyncs</code>	5s	并发执行的 Replication Controller 的同步任务的数量
<code>NodeSyncPeriod</code>	5s	从 CloudProvider 处同步 Node 节点的周期
<code>NodeMonitorPeriod</code>	5s	Node 节点监控的周期
<code>ResourceQuotaSyncPeriod</code>	10s	对资源的配额使用情况进行同步的周期
<code>NamespaceSyncPeriod</code>	5min	Namespace 同步的周期
<code>PVClaimBinderSyncPeriod</code>	10s	对 PV（持久存储）和 PV 的申请进行同步的周期
<code>PodEvictionTimeout</code>	5min	在 Node 失败的情况下，其上的 Pod 多久后才被删除
<code>master</code>		Kubernetes API Server 的访问地址

从上述这些变量来看，`Controller Manager Server` 其实就是一个“超级调度中心”，它负责定期同步 Node 节点状态、资源使用配额信息、Replication Controller、Namespace、Pod 的 PV 绑定等信息，也包括执行诸如监控 Node 节点状态、清除失败的 Pod 容器记录等一系列定时任务。

在 `controller-manager.go` 里创建 `CMServer` 实例并把参数从命令行中传递到 `CMServer` 后，就调用它的 `func (s *CMServer) Run(_ []string)` 方法进入关键流程，这里首先创建一个 `Rest Client` 对象用于访问 `Kubernetes API Server` 提供的 API 服务：

```
kubeClient, err := client.New(kubeconfig)
if err != nil {
```

```
glog.Fatalf("Invalid API configuration: %v", err)
}
```

随后，创建一个 HTTP Server 以提供必要的性能分析（Performance Profile）和性能指标度量（Metrics）的 Rest 服务：

```
go func() {
    mux := http.NewServeMux()
    healthz.InstallHandler(mux)
    if s.EnableProfiling {
        mux.HandleFunc("/debug/pprof/", pprof.Index)
        mux.HandleFunc("/debug/pprof/profile", pprof.Profile)
        mux.HandleFunc("/debug/pprof/symbol", pprof.Symbol)
    }
    mux.Handle("/metrics", prometheus.Handler())

    server := &http.Server{
        Addr:    net.JoinHostPort(s.Address.String(),
strconv.Itoa(s.Port)),
        Handler: mux,
    }
    glog.Fatal(server.ListenAndServe())
}()
```

我们注意到性能分析的 Rest 路径是以/debug 开头的，表明是为了程序调试所用，事实上的确如此，这里的几个 Profile 选项都是针对当前 Go 进程的 Profile 数据，比如我们在 Master 节点上执行 curl 命令（地址为 <http://127.0.0.1:10252/debug/pprof/heap>）可以获取进程的当前堆栈信息，会输出如下信息：

```
heap profile: 4: 78112 [1109: 824584] @ heap/1048576
1: 32768 [1: 32768] @ 0x402612 0x75ab95 0x771419 0x771379 0x565f08 0x46133f
0x400d10 0x4155a3 0x43e711
1: 32768 [1: 32768] @ 0x408806 0x407968 0x97e591 0x9895aa 0x76099b 0xa2f400
0xa4e887 0x765dc4 0x557fbc 0x782fac 0x5fe5db 0x602ca7 0x462c92 0x400f06 0x415594
0x43e711
1: 12288 [1: 12288] @ 0x4199fc 0x7df75d 0x5b585c 0x5b4947 0x5b405a 0x5aa472
0x5aa2b7 0x5aa188 0x5ad0d3 0x46291e 0x43e711
1: 288 [1: 288] @ 0x415d6a 0x43276f 0x43510f 0x42fd37 0x4311f9 0x430ef5 0x43c136
```

其他还有 GC 回收、Symbol 查看、进程 30s 内的 CPU 利用率、协程的阻塞状态等 Profile 功能，输出的数据格式符合 google-perftools 这个工具的要求，因此可以做运行期的可视化 Profile，以便排查当前进程潜在的问题或性能瓶颈。

性能指标度量目前主要收集和统计 Kubernetes API Server 的 Rest API 的调用情况，执行 curl（<http://127.0.0.1:10252/metrics>），可以看到输出中包括大量类似下面的内容：

```
rest_client_request_latency_microseconds{url="http://centos-master:8080/api/v1/namespaces/default/endpoints/%3Cname%3E",verb="GET",quantile="0.5"} 1448
rest_client_request_latency_microseconds{url="http://centos-master:8080/api/v1/namespaces/default/endpoints/%3Cname%3E",verb="GET",quantile="0.9"} 1699
rest_client_request_latency_microseconds{url="http://centos-master:8080/api/v1/namespaces/default/endpoints/%3Cname%3E",verb="GET",quantile="0.99"} 2093
```

这些指标有助于协助发现 Controller Manager Server 在调度方面的性能瓶颈，因此可以理解为什么会被包括到进程代码中去。

接下来，启动流程进入到关键代码部分。在这里，启动进程分别创建如下控制器，这些控制器的主要目的是实现资源在 Kubernetes API Server 的注册表中的周期性同步工作：

- ◎ EndpointController 负责对注册表中的 Kubernetes Service 的 Endpoints 信息的同步工作；
- ◎ ReplicationManager 根据注册表中对 ReplicationController 的定义，完成 Pod 的复制或者移除，以确保复制数量的一致性；
- ◎ NodeController 则通过 CloudProvider 的接口完成 Node 实例的同步工作；
- ◎ servicecontroller 通过 CloudProvider 的接口完成云平台中的服务的同步工作，这些服务目前主要是外部的负载均衡服务；
- ◎ ResourceQuotaManager 负责资源配额使用情况的同步工作；
- ◎ NamespaceManager 负责 Namespace 的同步工作；
- ◎ PersistentVolumeClaimBinder 与 PersistentVolumeRecycler 分别完成 PersistentVolume 的绑定和回收工作；
- ◎ TokensController、ServiceAccountsController 分别完成 Kubernetes 服务的 Token、Account 的同步工作。

创建并启动完成上述的控制器以后，各个控制器就开始独立工作，Controller Manager Server 启动完毕。

### 6.3.2 关键代码分析

在 6.3.1 节对 kube-controller-manager 进程的启动过程进行了详细分析，我们发现这个进程的主要逻辑就是启动一系列的“控制器”。这里以 Kubernetes 里比较关键的 Pod 副本（Pod Replica）数量的控制实现过程为例，来分析完成这个任务的“控制器”——ReplicationManager 具体是如何工作的。

首先，我们来看看 ReplicationManager 结构体的定义：

```
type ReplicationManager struct {
    kubeClient client.Interface
    podControl PodControlInterface

    // An rc is temporarily suspended after creating/deleting these many replicas.
    // It resumes normal action after observing the watch events for them.
    burstReplicas int
    // To allow injection of syncReplicationController for testing.
    syncHandler func(rcKey string) error

    // podStoreSynced returns true if the pod store has been synced at least once.
    // Added as a member to the struct to allow injection for testing.
    podStoreSynced func() bool

    // A TTLCache of pod creates/deletes each rc expects to see
    expectations RCExpectationsManager
    // A store of controllers, populated by the rcController
    controllerStore cache.StoreToControllerLister
    // A store of pods, populated by the podController
    podStore cache.StoreToPodLister
    // Watches changes to all replication controllers
    rcController *framework.Controller
    // Watches changes to all pods
    podController *framework.Controller
    // Controllers that need to be updated
    queue *workqueue.Type
}
```

在上述结构体里，比较关键的几个属性如下。

- ◎ **kubeClient**：用来访问 Kubernetes API Server 的 Rest 客户端，这里用来访问注册表中定义的 ReplicationController 对象并操作 Pod。
- ◎ **podControl**：实现了 Pod 副本创建的函数，其实现类为 RealPodControl（位于 `kubernetes/pkg/controller/controller_utils.go`）。
- ◎ **syncHandler**：是 RC（ReplicationController）的同步实现方法，完成具体的 RC 同步逻辑（创建 Pod 副本时调用 PodControl 的相关方法），在代码中其被赋值为 ReplicationManager.syncReplicationController 方法。
- ◎ **expectations**：是 Pod 副本在创建、删除过程中的流控机制的重要组成部分。
- ◎ **controllerStore**：是一个具备本地缓存功能的通用的资源存储服务，这里存放 framework.Controller 运行过程中从 Kubernetes API Server 同步过来的资源数据，目的是减轻资源同步过程中对 Kubernetes API Server 造成的访问压力并提高资源同步的效率。
- ◎ **rcController**：framework.Controller 的一个实例，用来实现 RC 同步的任务调度逻辑。

- ◎ `framework.Controller`: 是 `kube-controller-manager` 里设计的用于资源对象同步逻辑的专用任务调度框架。
- ◎ `podStore`: 类似于 `controllerStore` 的作用, 用来存取和获取 Pod 资源对象。
- ◎ `podController`: 类似于 `rcController` 的作用, 用来实现 Pod 同步的任务调度逻辑。

理解了 `ReplicationManager` 结构体的重要参数及其作用之后, 我们来看 `controller.NewReplicationManager(kubeClient client.Interface, burstReplicas int) *ReplicationManager` 这个构造函数中的关键代码, 注意到这里通过调用 `framework.NewInformer()` 方法先后创建了用于 RC 同步及 Pod 同步的 `framework.Controller`。下面是 `framework.NewInformer()` 方法的源码:

```
func NewInformer(
    lw cache.ListerWatcher,
    objType runtime.Object,
    resyncPeriod time.Duration,
    h ResourceEventHandler,
) (cache.Store, *Controller) {
    clientState := cache.NewStore(DeletionHandlingMetaNamespaceKeyFunc)
    fifo := cache.NewDeltaFIFO(cache.MetaNamespaceKeyFunc, nil, clientState)
    cfg := &Config{
        Queue:          fifo,
        ListerWatcher:   lw,
        ObjectType:      objType,
        FullResyncPeriod: resyncPeriod,
        RetryOnError:    false,
        Process: func(obj interface{}) error {
            // from oldest to newest
            for _, d := range obj.(cache.Deltas) {
                switch d.Type {
                    case cache.Sync, cache.Added, cache.Updated:
                        if old, exists, err := clientState.Get(d.Object); err == nil
&& exists {
                            if err := clientState.Update(d.Object); err != nil {
                                return err
                            }
                            h.OnUpdate(old, d.Object)
                        } else {
                            if err := clientState.Add(d.Object); err != nil {
                                return err
                            }
                            h.OnAdd(d.Object)
                        }
                    case cache.Deleted:
                        if err := clientState.Delete(d.Object); err != nil {
                            return err
                        }
                }
            }
        },
    }
```

```

    }
    h.OnDelete(d.Object)
  }
  return nil
},
}
return clientState, New(cfg)
}

```

在上述代码中，`lw(ListerWatcher)`用来获取和监测资源对象的变化，而 `fifo` 则是一个 `DeltaFIFO` 的 `Queue`，用来存放变化的资源（需要同步的资源）。当 `Controller` 框架发现有变化的资源需要处理时，就会将新资源与本地缓存 `clientState` 中的资源进行对比，然后调用相应的资源处理函数 `ResourceEventHandler` 的方法，完成具体的处理逻辑。下面是针对 `RC` 的 `ResourceEventHandler` 的具体实现：

```

framework.ResourceEventHandlerFuncs{
  AddFunc: rm.enqueueController,
  UpdateFunc: func(old, cur interface{}) {
    oldRC := old.(*api.ReplicationController)
    curRC := cur.(*api.ReplicationController)
    if oldRC.Status.Replicas != curRC.Status.Replicas {
      glog.V(4).Infof("Observed updated replica count for rc: %v, %d->%d", curRC.Name, oldRC.Status.Replicas, curRC.Status.Replicas)
    }
    rm.enqueueController(cur)
  },
  DeleteFunc: rm.enqueueController,
}

```

在上述源码中，我们看到当 `RC` 里 `Pod` 的副本数量属性发生变化以后，`ResourceEventHandler` 就将此 `RC` 放入 `ReplicationManager` 的 `queue` 队列中等待处理，为什么没有在这个 `handler` 函数中直接处理而是先放入队列再异步处理呢？最主要的一个原因是 `Pod` 副本创建的过程比较耗时。`Controller` 框架把需要同步的 `RC` 对象放入 `queue` 以后，接下来是谁在“消费”这个队列呢？答案就在 `ReplicationManager` 的 `Run()`方法中：

```

func (rm *ReplicationManager) Run(workers int, stopCh <-chan struct{}) {
  defer util.HandleCrash()
  go rm.rcController.Run(stopCh)
  go rm.podController.Run(stopCh)
  for i := 0; i < workers; i++ {
    go util.Until(rm.worker, time.Second, stopCh)
  }
  <-stopCh
  glog.Infof("Shutting down RC Manager")
  rm.queue.ShutDown()
}

```



```
}
```

上述代码首先启动 `rcController` 与 `podController` 这两个 `Controller`，启动之后，这两个 `Controller` 就分别开始拉取 RC 与 Pod 的变动信息，随后又启动  $N$  个协程并发处理 RC 的队列，其中 `func Until (f func(), period time.Duration, stopCh <-chan struct{})` 方法的逻辑是按照指定的周期 `period` 执行方法 `f`。下面是 `ReplicationManager` 的 `worker` 方法的源码，负责从 RC 队列中拉取 RC 并调用 `rm` 的 `syncHandler` 方法完成具体处理：

```
func (rm *ReplicationManager) worker() {
    for {
        func() {
            key, quit := rm.queue.Get()
            if quit {
                return
            }
            defer rm.queue.Done(key)
            err := rm.syncHandler(key.(string))
            if err != nil {
                glog.Errorf("Error syncing replication controller: %v", err)
            }
        }()
    }
}
```

从 `ReplicationManager` 的构造函数中我们得知：`syncHandler` 在这里其实是 `func (rm *ReplicationManager) syncReplicationController(key string) error` 方法。下面是该方法的源码：

```
func (rm *ReplicationManager) syncReplicationController(key string) error {
    startTime := time.Now()
    defer func() {
        glog.V(4).Infof("Finished syncing controller %q (%v) ", key, time.
Now().Sub(startTime))
    }()

    obj, exists, err := rm.controllerStore.Store.GetByKey(key)
    if !exists {
        glog.Infof("Replication Controller has been deleted %v", key)
        rm.expectations.DeleteExpectations(key)
        return nil
    }
    if err != nil {
        glog.Infof("Unable to retrieve rc %v from store: %v", key, err)
        rm.queue.Add(key)
        return err
    }
    controller := *obj.(*api.ReplicationController)
    if !rm.podStoreSynced() {
```

```

        // Sleep so we give the pod reflector goroutine a chance to run.
        time.Sleep(PodStoreSyncedPollPeriod)
        glog.Infof("Waiting for pods controller to sync, requeuing rc %v",
controller.Name)
        rm.enqueueController(&controller)
        return nil
    }

    rcNeedsSync := rm.expectations.SatisfiedExpectations(&controller)
    podList, err := rm.podStore.Pods(controller.Namespace).List(labels.Set
(controller.Spec.Selector).AsSelector())
    if err != nil {
        glog.Errorf("Error getting pods for rc %q: %v", key, err)
        rm.queue.Add(key)
        return err
    }

    filteredPods := filterActivePods(podList.Items)
    if rcNeedsSync {
        rm.manageReplicas(filteredPods, &controller)
    }

    if err := updateReplicaCount(rm.kubeClient.ReplicationControllers(controller.
Namespace), controller, len(filteredPods)); err != nil {
        rm.enqueueController(&controller)
    }
    return nil
}

```

在上述代码里有一个重要的流控变量 `rcNeedsSync`。为了限流，在 RC 同步逻辑的过程中，一个 RC 每次最多执行  $N$  个 Pod 的创建、删除，如果某个 RC 的同步过程涉及的 Pod 副本数量超过 `burstReplicas` 这个阈值，就会采用 `RCEExpectations` 机制进行限流。`RCEExpectations` 对象可以理解为一个简单的规则：即在限定的时间内执行  $N$  次操作，每次操作都使计数器减一，计数器为零表示  $N$  个操作已经完成，可以进行下一批次的操作了。

Kubernetes 为什么会设计这样一个流程控制机制？其实答案很简单——为了公平。因为谷歌的开发 Kubernetes 的资深大牛们早已预见到某个 RC 的 Pod 副本一次扩容至 100 倍的极端情况可能真实发生，如果没有流控机制，则这个巨无霸的 RC 同步操作会导致其他众多“散户”崩溃！这绝对不是谷歌的理念。

接着看上述代码里所调用的 `ReplicationManager` 的 `manageReplicas` 方法，这是 RC 同步的具体逻辑实现，此方法采用了并发调用的方式执行批量的 Pod 副本操作任务，相关代码如下：

```

wait := sync.WaitGroup{}
wait.Add(diff)

```

```

        glog.V(2).Infof("Too few %q/%q replicas, need %d, creating %d",
controller.Namespace, controller.Name, controller.Spec.Replicas, diff)
        for i := 0; i < diff; i++ {
            go func() {
                defer wait.Done()
                if err := rm.podControl.createReplica(controller.Namespace,
controller); err != nil {
                    glog.V(2).Infof("Failed creation, decrementing expectations for
controller %q/%q", controller.Namespace, controller.Name)
                    rm.expectations.CreationObserved(controller)
                    util.HandleError(err)
                }
            }()
        }
        wait.Wait()

```

追踪至此，我们才看到创建 Pod 副本的真正代码在 `PodControl.createReplica()` 方法里，而此方法的具体实现方法则是 `RealPodControl.createReplica()`，位于 `controller_utils.go` 里。通过分析该方法，我们可以知道创建 Pod 副本的过程就是创建一个 Pod 资源对象，并把 RC 中定义的 Pod 模板赋值给该 Pod 对象，并且 Pod 的名字将 RC 的名字作为前缀，最后调用 `Kubernetes Client` 将 Pod 对象通过 `Kubernetes API Server` 写入后端的 `etcd` 存储中。

在本节最后，我们来分析一下 `Controller` 框架中如何实现资源对象的查询和监听逻辑并且在资源发生变动时回调 `Controller.Config` 对象中的 `Process` 方法：`func(obj interface{})`，最终完成整个 `Controller` 框架的闭环过程。

首先，在 `Controller` 框架中构建了 `Reflector` 对象以实现资源对象的查询和监听逻辑，它的源码位于 `pkg/client/cache/reflector.go` 中，我们看一下这个对象的数据结构就基本明白了其工作原理：

```

// Reflector watches a specified resource and causes all changes to be reflected
in the given store.
type Reflector struct {
    // The type of object we expect to place in the store.
    expectedType reflect.Type
    // The destination to sync up with the watch source
    store Store
    // listerWatcher is used to perform lists and watches.
    listerWatcher ListerWatcher
    // period controls timing between one watch ending and
    // the beginning of the next one.
    period time.Duration
    resyncPeriod time.Duration
    // lastSyncResourceVersion is the resource version token last
    // observed when doing a sync with the underlying store

```

```
// it is thread safe, but not synchronized with the underlying store
lastSyncResourceVersion string
// lastSyncResourceVersionMutex guards read/write access to
lastSyncResourceVersion
lastSyncResourceVersionMutex sync.RWMutex
}
```

核心思路就是通过 `listerWatcher` 去获取资源列表并监听资源的变化，然后存储到 `store` 中。这里你可能有个疑问，这个 `store` 究竟是哪个对象？是 `ReplicationManager` 里的 `controllerStore` 还是 `framework.NewInformer()` 方法里创建的 `fifo` 队列？

下面的两段来自 `pkg/controller/framework/controller.go` 的代码会告诉我们答案。

首先是来自 `Controller` 的 `run` 方法 `func (c *Controller) Run(stopCh <-chan struct{})` 的代码片段：

```
r := cache.NewReflector(
    c.config.ListerWatcher,
    c.config.ObjectType,
    c.config.Queue,
    c.config.FullResyncPeriod,
)
```

然后是来自 `Controller` 的 `NewInformer` 方法 `func NewInformer(lw cache.ListerWatcher, objType runtime.Object, resyncPeriod time.Duration, h ResourceEventHandler,) (cache.Store, *Controller)` 中的代码片段：

```
cfg := &Config{
    Queue:           fifo,
    ListerWatcher:   lw,
    ObjectType:      objType,
    FullResyncPeriod: resyncPeriod,
    RetryOnError:    false,
```

分析上述代码，我们发现 `Reflector` 中的 `store` 其实是引用 `Controller.Config` 里的 `Queue` 属性，即 `fifo` 队列，而非 `ReplicationManager` 里的 `controllerStore`。我们费了这么大的劲，才弄明白这个简单的问题，这告诉我们一个事实：编程中有良好的命名规则很重要。

下面这段代码是 `Controller` 从队列 `Queue` 中拉取资源对象并且交给 `Controller.Config` 对象中的 `Process` 方法 `func(obj interface{})` 进行处理，从而最终完成了整个 `Controller` 框架的闭环过程。

```
func (c *Controller) processLoop() {
    for {
        obj := c.config.Queue.Pop()
        err := c.config.Process(obj)
        if err != nil {
```

```

        if c.config.RetryOnError {
            // This is the safe way to re-enqueue.
            c.config.Queue.AddIfNotPresent(obj)
        }
    }
}
}

```

至于上述过程的调用则是在 Controller 启动（Run 方法）的最后一步里，Controller 框架定时每秒调用一次上述函数，代码如下：

```
util.Until(c.processLoop, time.Second, stopCh)
```

最后，给读者留一个源码解读的问题，即 ReplicationManager 里除了 RC Controller，又构造了一个用于 Pod 的 Controller，它的逻辑具体是怎样实现的？它与 RC Controller 是怎样交互的？

### 6.3.3 设计总结

相对于之前的 Kubernetes API Server 设计来说，Kubernetes Controller Server 的设计没有那么复杂，而且精彩依旧。不愧是大师的作品，Controller Framework 精巧细致的设计使得整个进程中各种资源对象的同步逻辑在代码实现方面保持了高度一致性与简捷性。此外，在关键资源 RC（Replication Controller）的同步逻辑中所采用的流控机制也简单、高效。

本节我们针对 Kubernetes Controller Server 中的精华部分——Controller Framework 的设计做一个整理分析。首先，framework.Controller 内部维护一个 Config 对象，保留了一个标准的消息、事件分发系统的三要素。

- ◎ 生产者：cache.ListerWatch。
- ◎ 队列：cache.cacheStore(Queue)。
- ◎ 消费者：用回调函数来模拟(framework.ResourceEventHandlerFuncs)。

由于生产者的逻辑比较复杂，在这个系统中也有其特殊性，即拉取资源并监控资源的变化，由此产生了真正的待处理任务，所以又设计了一个 ListerWatcher 接口，将底层的复杂逻辑“框架化”，放入 cache.Reflector 中，使用者只要简单地实现 ListerWatcher 接口的 ListFunc 与 WatchFunc 即可。另外，cache.Reflector 也是独立于 Controller Framework 的一个组件，隶属于 cache 包，它的功能是将任意资源对象拉取到本地缓存中并监控资源的变化，保持本地缓存的同步，其目标是减轻对 Kubernetes API Server 的请求压力。

图 6.4 给出了 Controller Framework 的整体架构设计图。

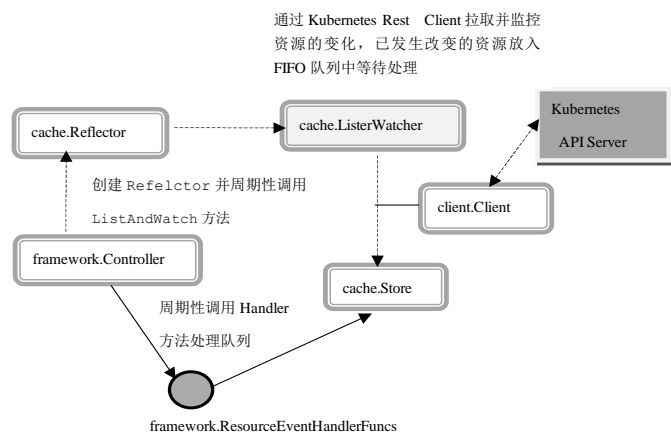


图 6.4 Controller Framework 的整体架构设计图

Kubernetes Controller Server 中所有涉及同步的资源都采用了 Controller Framework 框架来进行驱动。如图 6.5 所示为 Kubernetes Controller Server 的整体设计示意图。

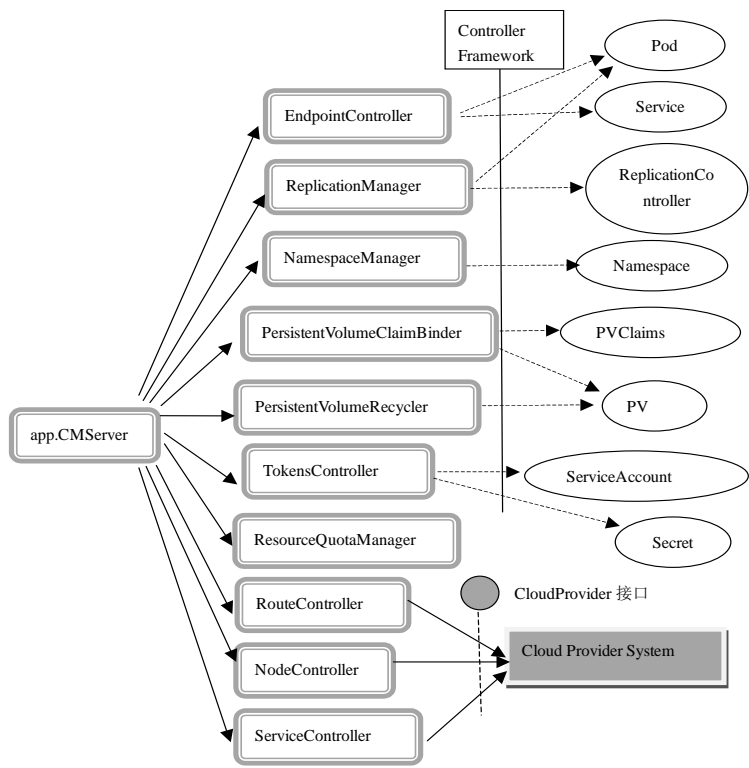


图 6.5 Kubernetes Controller Server 的整体设计示意图

可以看出，除了 Node、Route、Cloud Service 这三个资源依赖于 Kubernetes 所处的云计算环境，只能通过 CloudProvider 接口所提供的 API 来完成资源同步，其他资源都采用了 Controller Framework 框架来进行资源同步。图中的虚线箭头表示针对目标资源创建了一个 framework.Controller 对象，其中的某些资源如 RC、PV、Tokens 的同步过程需要获取并监听其他与之相关联的资源对象。这里只有 ResourceQuota 资源比较另类，它没有采用 Controller Framework，一个原因是 ResourceQuota 涉及很多资源对象，不大好应用 framework.Controller，另外一个原因可能是写 ResourceQuotaManager 的专家拥有比较浪漫的情怀，看看下面这段 Kubernetes 中最优美的代码吧：

```
func (rm *ResourceQuotaManager) Run(period time.Duration) {
    rm.syncTime = time.Tick(period)
    go util.Forever(func() { rm.synchronize() }, period)
}
```

核心代码翻译过来就是这个意思：从此他们过上了幸福的生活，一去不复返了！

## 6.4 kube-scheduler 进程源码分析

Kubernetes Scheduler Server 是由 kube-scheduler 进程实现的，它运行在 Kubernetes 的管理节点——Master 上并主要负责完成从 Pod 到 Node 的调度过程。Kubernetes Scheduler Server 跟踪 Kubernetes 集群中所有 Node 的资源利用情况，并采取合适的调度策略，确保调度的均衡性，避免集群中的某些节点“过载”。从某种意义上来说，Kubernetes Scheduler Server 也是 Kubernetes 集群的“大脑”。

谷歌作为公有云的重要供应商，积累了很多经验并且了解客户的需求。在谷歌看来，客户并不真正关心他们的服务究竟运行在哪台机器上，他们最关心服务的可靠性，希望发生故障后能自动恢复。遵循这一指导思想，Kubernetes Scheduler Server 实现了“完全市场经济”的调度原则并彻底抛弃了传统意义上的“计划经济”。

下面我们分别对其启动过程、关键代码分析及设计总结等方面进行深入分析和讲解。

### 6.4.1 进程启动过程

kube-scheduler 进程的入口类源码位置如下：

```
github.com/GoogleCloudPlatform/kubernetes/plugin/cmd/kube-scheduler/scheduler.go.
```

入口 main() 函数的逻辑如下：

```
func main() {
    runtime.GOMAXPROCS(runtime.NumCPU())
    s := app.NewSchedulerServer()
    s.AddFlags(pflag.CommandLine)
    util.InitFlags()
    util.InitLogs()
    defer util.FlushLogs()
    verflag.PrintAndExitIfRequested()
    s.Run(pflag.CommandLine.Args())
}
```

对上述代码的风格和逻辑我们再熟悉不过了：创建一个 `SchedulerServer` 对象，将命令行参数传入，并且进入 `SchedulerServer` 的 `Run` 方法，无限循环下去。

按照惯例，我们首先看看 `SchedulerServer` 的数据结构（`app/server.go`），下面是其定义：

```
type SchedulerServer struct {
    Port          int
    Address        util.IP
    AlgorithmProvider string
    PolicyConfigFile string
    EnableProfiling bool
    Master         string
    Kubeconfig     string
}
```

这里的关键属性有以下两个。

- ◎ **AlgorithmProvider**：对应参数 `algorithm-provider`，是 `AlgorithmProviderConfig` 的名称。
- ◎ **PolicyConfigFile**：用来加载调度策略文件。

从代码上来看这两个参数的作用其实是一样的，都是加载一组调度规则，这组调度规则要么在程序里定义为一个 `AlgorithmProviderConfig`，要么保存到文件中。下面的源码清楚地解释了这个过程：

```
func (s *SchedulerServer) createConfig(configFactory *factory.ConfigFactory)
(*scheduler.Config, error) {
    var policy schedulerapi.Policy
    var configData []byte

    if _, err := os.Stat(s.PolicyConfigFile); err == nil {
        configData, err = ioutil.ReadFile(s.PolicyConfigFile)
        if err != nil {
            return nil, fmt.Errorf("Unable to read policy config: %v", err)
        }
        err = latestschedulerapi.Codec.DecodeInto(configData, &policy)
        if err != nil {
            return nil, fmt.Errorf("Invalid configuration: %v", err)
        }
    }
}
```



```

    }

    return configFactory.CreateFromConfig(policy)
}

// if the config file isn't provided, use the specified (or default) provider
// check of algorithm provider is registered and fail fast
_, err := factory.GetAlgorithmProvider(s.AlgorithmProvider)
if err != nil {
    return nil, err
}

return configFactory.CreateFromProvider(s.AlgorithmProvider)
}

```

创建了 `SchedulerServer` 结构体实例后，调用此实例的方法 `func (s *APIServer) Run(_[]string)`，进入关键流程。首先，创建一个 `Rest Client` 对象用于访问 Kubernetes API Server 提供的 API 服务：

```

    kubeClient, err := client.New(kubeconfig)
    if err != nil {
        glog.Fatalf("Invalid API configuration: %v", err)
    }

```

随后，创建一个 `HTTP Server` 以提供必要的性能分析（Performance Profile）和性能指标度量（Metrics）的 Rest 服务：

```

go func() {
    mux := http.NewServeMux()
    healthz.InstallHandler(mux)
    if s.EnableProfiling {
        mux.HandleFunc("/debug/pprof/", pprof.Index)
        mux.HandleFunc("/debug/pprof/profile", pprof.Profile)
        mux.HandleFunc("/debug/pprof/symbol", pprof.Symbol)
    }
    mux.Handle("/metrics", prometheus.Handler())

    server := &http.Server{
        Addr:    net.JoinHostPort(s.Address.String(), strconv.Itoa(s.Port)),
        Handler: mux,
    }
    glog.Fatal(server.ListenAndServe())
}()

```

接下来，启动程序构造了 `ConfigFactory`，这个结构体包括了创建一个 `Scheduler` 所需的必要属性。

- ◎ `PodQueue`：需要调度的 Pod 队列。
- ◎ `BindPodsRateLimiter`：调度过程中限制 Pod 绑定速度的限速器。

- ◎ **modeler**：这是用于优化 Pod 调度过程而设计的一个特殊对象，用于“预测未来”。一个 Pod 被计划调度到机器 A 的事实被称为 **assumed** 调度，即假定调度，这些调度安排被保存到特定队列里，此时调度过程是能看到这个预安排的，因而会影响到其他 Pod 的调度。
- ◎ **PodLister**：负责拉取已经调度过的，以及被假定调度过的 Pod 列表。
- ◎ **NodeLister**：负责拉取 Node 节点（Minion）列表。
- ◎ **ServiceLister**：负责拉取 Kubernetes 服务列表。
- ◎ **ScheduledPodLister**、**scheduledPodPopulator**：Controller 框架创建过程中返回的 Store 对象与 controller 对象，负责定期从 Kubernetes API Server 上拉取已经调度好的 Pod 列表，并将这些 Pod 从 modeler 的假定调度过的队列中删除。

在构造 ConfigFactory 的方法 `factory.NewConfigFactory(kubeClient)` 中，我们看到下面这段代码：

```
c.ScheduledPodLister.Store, c.scheduledPodPopulator = framework.NewInformer(
    c.createAssignedPodLW(),
    &api.Pod{},
    0,
    framework.ResourceEventHandlerFuncs{
        AddFunc: func(obj interface{}) {
            if pod, ok := obj.(*api.Pod); ok {
                c.modeler.LockedAction(func() {
                    c.modeler.ForgetPod(pod)
                })
            }
        },
        DeleteFunc: func(obj interface{}) {
            c.modeler.LockedAction(func() {
                switch t := obj.(type) {
                case *api.Pod:
                    c.modeler.ForgetPod(t)
                case cache.DeletedFinalStateUnknown:
                    c.modeler.ForgetPodByKey(t.Key)
                }
            })
        },
    },
)
```

这里沿用了之前看到的 controller framework 的身影，上述 Controller 实例所做的事情是获取并监听已经调度的 Pod 列表，并将这些 Pod 列表从 modeler 中的“assumed”队列中删除。

接下来，启动进程用上述创建好的 `ConfigFactory` 对象作为参数来调用 `SchedulerServer` 的 `createConfig` 方法，创建一个 `Scheduler.Config` 对象，而此段代码的关键逻辑则集中在 `ConfigFactory` 的 `CreateFromKeys` 这个函数里，其主要步骤如下。

(1) 创建一个与 Pod 相关的 `Reflector` 对象并定期执行，该 `Reflector` 负责查询并监测等待调度的 Pod 列表，即还没有分配主机的 Pod (`Unsigned Pod`)，然后把它们放入 `ConfigFactory` 的 `PodQueue` 中等待调度。相关代码为：`cache.NewReflector(f.createUnassignedPodLW(), &api.Pod{}, f.PodQueue, 0).RunUntil(f.StopEverything)`。

(2) 启动 `ConfigFactory` 的 `scheduledPodPopulator Controller` 对象，负责定期从 `Kubernetes API Server` 上拉取已经调度好的 Pod 列表，并将这些 Pod 从 `modeler` 中的假定 (`assumed`) 调度过的队列中删除。相关代码为：`go f.scheduledPodPopulator.Run(f.StopEverything)`。

(3) 创建一个 Node 相关的 `Reflector` 对象并定期执行，该 `Reflector` 负责查询并监测可用的 Node 列表(可用意味着 Node 的 `spec.unschedulable` 属性为 `false`)，这些 Node 被放入 `ConfigFactory` 的 `NodeLister.Store` 里。相关代码为：`cache.NewReflector(f.createMinionLW(), &api.Node{}, f.NodeLister.Store, 0).RunUntil(f.StopEverything)`。

(4) 创建一个 Service 相关的 `Reflector` 对象并定期执行，该 `Reflector` 负责查询并监测已定义的 Service 列表，并放入 `ConfigFactory` 的 `ServiceLister.Store` 里。这个过程的目的是 `Scheduler` 需要知道一个 Service 当前所创建的所有 Pod，以便能正确地进行调度。相关代码为：`cache.NewReflector(f.createServiceLW(), &api.Service{}, f.ServiceLister.Store, 0).RunUntil(f.StopEverything)`。

(5) 创建一个实现了 `algorithm.ScheduleAlgorithm` 接口的对象 `genericScheduler`，它负责完成从 Pod 到 Node 的具体调度工作，调度完成的 Pod 放入 `ConfigFactory` 的 `PodLister` 里。相关代码为 `algo := scheduler.NewGenericScheduler(predicateFuncs, priorityConfigs, f.PodLister, r)`。

(6) 最后一步，使用之前的这些信息创建 `Scheduler.Config` 对象并返回。

从上面的分析我们看出，其实在创建 `Scheduler.Config` 的过程中已经完成了 `Kubernetes Scheduler Server` 进程中的很多启动工作，于是整个进程的启动过程的最后一步简单明了：使用刚刚创建好的 `Config` 对象来构造一个 `Scheduler` 对象并启动运行。即下面的两行代码：

```
sched := scheduler.New(config)
sched.Run()
```

而 `Scheduler` 的 `Run` 方法就是不停地执行 `scheduleOne` 方法：

```
go util.Until(s.scheduleOne, 0, s.config.StopEverything)
```

`scheduleOne` 方法的逻辑也比较清晰，即获取下一个待调度的 Pod，然后交给 `genericScheduler` 进行调度（完成 Pod 到某个 Node 的绑定过程），调度成功以后通知 `Modeler`。这个过程同时增加了限流和性能指标的逻辑。

### 6.4.2 关键代码分析

在 6.4.1 节对 kube-scheduler 进程的启动过程进行详细分析后，我们大致明白了 Kubernetes Scheduler Server 的工作流程，但由于代码中涉及多个 Pod 队列和 Pod 状态切换逻辑，因此这里有必要对这个问题进行详细分析，以弄清在整个调度过程中 Pod 的“来龙去脉”。首先，我们知道 ConfigFactory 里的 PodQueue 是“待调度的 Pod 队列”，这个过程是通过无限循环执行一个 Reflector 来从 Kubernetes API Server 上获取待调度的 Pod 列表并填充到队列中实现的，因为 Reflector 框架已经实现了通用的代码，所以到了 Kubernetes Scheduler Server 这里，通过一行代码就能完成这个复杂的过程：

```
cache.NewReflector(f.createUnassignedPodLW(), &api.Pod{}, f.PodQueue, 0).
RunUntil(f.StopEverything)
```

上述代码中的 createUnassignedPodLW 是查询和监测 spec.nodeName 为空的 Pod 列表，此外，我们注意到 scheduler.Config 里提供了 NextPod 这个函数指针来从上述队列中消费一个元素，下面是相关代码片段（来自 ConfigFactory 的 CreateFromKeys 方法中创建 scheduler.Config 的代码）：

```
NextPod: func() *api.Pod {
    pod := f.PodQueue.Pop().(*api.Pod)
    glog.V(2).Infof("About to try and schedule pod %v", pod.Name)
    return pod
},
```

然后，这个 PodQueue 是怎样被消费的呢？就在之前提到的 Scheduler.scheduleOne 的方法里，每次调用 NextPod 方法会获取一个可用的 Pod，然后交给 genericScheduler 进行调度，下面是相关代码片段（省略了其他代码）：

```
pod := s.config.NextPod()
if s.config.BindPodsRateLimiter != nil {
    s.config.BindPodsRateLimiter.Accept()
}
dest, err := s.config.Algorithm.Schedule(pod, s.config.MinionLister)
```

genericScheduler.Schedule 方法只是给出该 Pod 调度到的目标 Node，如果调度成功，则设置该 Pod 的 spec.nodeName 为目标 Node，然后通过 HTTP Rest 调用写入 Kubernetes API Server 里完成 Pod 的 Binding 操作，最后通知 ConfigFactory 的 modeler（具体实例对应 scheduler.SimpleModeler），将此 Pod 放入 Assumed Pod 队列，下面是相关代码片段：

```
s.config.Modeler.LockedAction(func() {
    bindingStart := time.Now()
    err := s.config.Binder.Bind(b)

    metrics.BindingLatency.Observe(metrics.SinceInMicroseconds(bindingStart))
```

```

        s.config.Recorder.Eventf(pod, "scheduled", "Successfully assigned %v to
        %v", pod.Name, dest)
        // tell the model to assume that this binding took effect.
        assumed := *pod
        assumed.Spec.NodeName = dest
        s.config.Modeler.AssumePod(&assumed)
    })

```

当 Pod 执行 Bind 操作成功以后，Kubernetes API Server 上 Pod 已经满足“已调度”的条件，因为 spec.nodeName 已经被设置为目标 Node 地址，此时 ConfigFactory 的 scheduledPodPopulator 这个 Controller 就会监听至此变化，将此 Pod 从 modeler 中的 Assumed 队列中删除，下面是相关代码片段：

```

        framework.ResourceEventHandlerFuncs{
            AddFunc: func(obj interface{}) {
                if pod, ok := obj.(*api.Pod); ok {
                    c.modeler.LockedAction(func() {
                        c.modeler.ForgetPod(pod)
                    })
                }
            },
            .....,
        },
    },

```

谷歌的大神在源码中说明 Modeler 的存在是为了调度的优化，那么这个优化具体体现在哪里呢？由于 Rest Watch API 存在延时，当前已经调度好的 Pod 很可能还未被通知给 Scheduler，于是大神灵光一闪：为每个刚刚调度完成的 Pod 发放一个“暂住证”，安排“暂住”到“Assumed”队列里，然后设计一个获取当前“已调度”的 Pod 队列的新方法，该方法合并 Assumed 队列与 Watch 缓存队列，这样一来，就得到了最佳答案。如果你打算看看这段代码，那么它就在 SimpleModeler 的 listPods 方法里，至此，你若也完全明白了 c.PodLister = modeler.PodLister() 这句简单却又深奥的代码，那么恭喜你，你离大神的距离又缩短了一个厘米。

接下来，我们深入分析 Pod 调度中所用到的流控技术，缘起于下面这段代码：

```

if s.config.BindPodsRateLimiter != nil {
    s.config.BindPodsRateLimiter.Accept()
}

```

上述代码中的 BindPodsRateLimiter 采用了开源项目 juju 的一个子项目 ratelimit，项目地址为 <https://github.com/juju/ratelimit>，它实现了一个高效的基于经典令牌桶（Token Bucket）的流控算法。如图 6.6 所示是经典令牌桶流控算法的原理示意图。

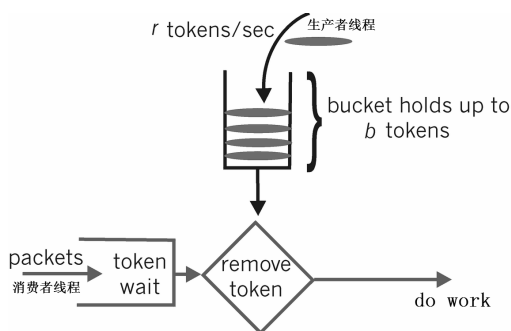


图 6.6 经典令牌桶流控算法的原理示意图

简单地说，控制线程以固定速率向一个固定容量的桶（Bucket）中投放令牌（Token），消费者线程则等待并获取到一个令牌后才能继续接下来的任务，否则需要等待可用令牌的到来。具体说来，假如用户配置的平均限流速率为  $r$ ，则每隔  $1/r$  s 就会有一个令牌被加入桶中，而令牌桶最多可以存储  $b$  个令牌，如果令牌到达时令牌桶已经满了，那么这个令牌会被丢弃。从长期运行结果来看，消费者的处理速率被限制成常量  $r$ 。令牌桶流控算法除了能够限制平均处理速度，还允许某种程度的突发速率。

juju 的 `ratelimit` 模块通过下面的 API 提供了构造一个令牌桶的简单做法，其中，`rate` 参数表示每秒填充到桶里的令牌数量，`capacity` 则是桶的容量：

```
func NewBucketWithRate(rate float64, capacity int64) *Bucket
```

我们回头再看看 Kubernetes Scheduler Server 中 `BindPodsRateLimiter` 的赋值代码：`c.BindPodsRateLimiter = util.NewTokenBucketRateLimiter(BindPodsQps, BindPodsBurst)`，跟踪进去，发现它就是调用了刚才所提到的 juju 函数 `limiter := ratelimit.NewBucketWithRate(float64(qps), int64(burst))`，其中 `qps` 目前为常量 15，而 `burst` 为 20，目前在 Kubernetes 1.0 版本中还没有提供命令行参数来配置此变量，会在未来的版本中实现。

最后，我们一起深入分析 Kubernetes Scheduler Server 中关于 Pod 调度的细节。首先，我们需要理解启动过程中 `SchedulerServer` 加载调度策略相关配置的这段代码：

```
predicateFuncs, err := getFitPredicateFunctions(predicateKeys, pluginArgs)
priorityConfigs, err := getPriorityFunctionConfigs(priorityKeys, pluginArgs)
algo := scheduler.NewGenericScheduler(predicateFuncs, priorityConfigs, f.
PodLister, r)
```

这里加载了两组策略，其中 `predicateFuncs` 是一个 Map，key 为 `FitPredicate` 的名称，value 为对应的 `algorithm.FitPredicate` 函数，它表明一个候选的 Node 是否满足当前 Pod 的调度要求，`FitPredicate` 函数的具体定义如下：

```
type FitPredicate func(pod *api.Pod, existingPods []*api.Pod, node string) (bool,
error)
```

FitPredicate 是 Pod 调度过程中必须满足的规则，只有顺利通过由所有 FitPredicate 组成的这道封锁线，一个 Node 才能拿到主会场的“入场券”，成为一个合格的“候选人”，等待下一步“评审”。目前系统提供的具体的 FitPredicate 实现都在 predicates.go 里，系统默认加载注册 FitPredicate 的地方在 defaultPredicates 方法里。

当有一组 Node 通过筛查成为“候选人”之后，需要有一种办法来选择“最优”的 Node，这就是接下来我们要介绍的 PriorityConfigs 所要做的事情了。PriorityConfigs 是一个数组，类型为 algorithm.PriorityConfig，PriorityConfig 包括一个 PriorityFunction 函数，用来计算并给出一组 Node 的优先级，下面是相关代码：

```
type PriorityConfig struct {
    Function PriorityFunction
    Weight    int
}
type PriorityFunction func(pod *api.Pod, podLister PodLister, minionLister
MinionLister) (HostPriorityList, error)
type HostPriorityList []HostPriority
func (h HostPriorityList) Len() int {
    return len(h)
}
func (h HostPriorityList) Less(i, j int) bool {
    if h[i].Score == h[j].Score {
        return h[i].Host < h[j].Host
    }
    return h[i].Score < h[j].Score
}
```

如果看到这里还是不太明白它的用途，那么认真读一读下面这段来自 genericScheduler 的计算候选节点优先级的 PrioritizeNodes 方法，你就能顿悟了：一个候选节点的优先级总分是所有评委老师(PriorityConfig)一起给出的“加权总分”，评委老师越是德高望重(PriorityConfig.Weight 越大)，他的评分影响力就越大：

```
combinedScores := map[string]int{}
for _, priorityConfig := range priorityConfigs {
    weight := priorityConfig.Weight
    // skip the priority function if the weight is specified as 0
    if weight == 0 {
        continue
    }
    priorityFunc := priorityConfig.Function
    prioritizedList, err := priorityFunc(pod, podLister, minionLister)
    if err != nil {
        return algorithm.HostPriorityList{}, err
    }
    for _, hostEntry := range prioritizedList {
```

```
        combinedScores[hostEntry.Host] += hostEntry.Score * weight
    }
}
for host, score := range combinedScores {
    glog.V(10).Infof("Host %s Score %d", host, score)
    result = append(result, algorithm.HostPriority{Host: host, Score: score})
}
return result, nil
```

接下来，我们看看系统初始化加载的默认的 Predicate 与 Priorities 有哪些，通过追踪代码，我们发现默认加载的代码位于 `plugin/pkg/scheduler/algorithmprovider/default/default.go` 的 `init` 函数里：

```
func init() {
    factory.RegisterAlgorithmProvider(factory.DefaultProvider, defaultPredicates(),
    defaultPriorities())
    // EqualPriority is a prioritizer function that gives an equal weight of one
    to all minions
    // Register the priority function so that its available
    // but do not include it as part of the default priorities
    factory.RegisterPriorityFunction("EqualPriority", scheduler.EqualPriority, 1)
}
```

跟踪进去后，我们看到系统默认加载的 predicates 有如下几种。

- ◎ PodFitsResources。
- ◎ MatchNodeSelector。
- ◎ HostName。

而默认加载的 priorities 则有如下几种。

- ◎ LeastRequestedPriority。
- ◎ BalancedResourceAllocation。
- ◎ ServiceSpreadingPriority。

从上述这些信息来看，Kubernetes 默认的调度指导原则是尽量均匀分布 Node 到不同的 Node 上，并且确保各个 Node 上的资源利用率基本保持一致，也就是说如果你有 100 台机器，则可能每个机器都被调度到，而不是只有其中的 20% 被调度到，哪怕每台机器都只利用了不到 10% 的资源，这不正是所谓的“韩信点兵，多多益善”么？

接下来我们以服务亲和性这个默认没有加载的 Predicate 为例，看看 Kubernetes 是如何通过 Policy 文件注册加载它的。下面是我们定义的一个 Policy 文件：

```
{
  "kind" : "Policy",
  "version" : "v1",
```



```

    "predicates" : [
        .....
        {"name" : "RegionZoneAffinity", "argument" : {"serviceAffinity" :
{"labels" : [{"region", "zone"]}}}
    ],
    "priorities" : [
        .....
        {"name" : "RackSpread", "weight" : 1, "argument" : {"serviceAnti
Affinity" : {"label" : "rack"}}}
    ]
}

```

首先，这个文件被映射成 `api.Policy` 对象（`plugin/pkg/scheduler/api/types.go`）。下面是其结构体定义：

```

type Policy struct {
    api.TypeMeta `json: ",inline"`
    // Holds the information to configure the fit predicate functions
    Predicates []PredicatePolicy `json: "predicates"`
    // Holds the information to configure the priority functions
    Priorities []PriorityPolicy `json: "priorities"`
}

```

我们看到 `policy` 文件中的 `predicates` 部分被映射为 `PredicatePolicy` 数组：

```

type PredicatePolicy struct {
    Name string `json: "name"`
    Argument *PredicateArgument `json: "argument"`
}

```

而 `PredicateArgument` 的定义如下，包括服务亲和性的相关属性 `ServiceAffinity`：

```

type PredicateArgument struct {
    ServiceAffinity *ServiceAffinity `json: "serviceAffinity"`
    LabelsPresence *LabelsPresence `json: "labelsPresence"`
}

```

策略文件被映射为 `api.Policy` 对象后，`PredicatePolicy` 部分的处理逻辑则交给下面的函数进行处理（`plugin/pkg/scheduler/factory/plugin.go`）：

```

func RegisterCustomFitPredicate(policy schedulerapi.PredicatePolicy) string {
    var predicateFactory FitPredicateFactory
    var ok bool
    validatePredicateOrDie(policy)
    // generate the predicate function, if a custom type is requested
    if policy.Argument != nil {
        if policy.Argument.ServiceAffinity != nil {
            predicateFactory = func(args PluginFactoryArgs) algorithm.
FitPredicate {
                return predicates.NewServiceAffinityPredicate(

```

```

        args.PodLister,
        args.ServiceLister,
        args.NodeInfo,
        policy.Argument.ServiceAffinity.Labels,
    )
}
} else if policy.Argument.LabelsPresence != nil {
    predicateFactory = func(args PluginFactoryArgs) algorithm.
FitPredicate {
    return predicates.NewNodeLabelPredicate(
        args.NodeInfo,
        policy.Argument.LabelsPresence.Labels,
        policy.Argument.LabelsPresence.Presence,
    )
}
}
}

```

在上面的代码中,当 `ServiceAffinity` 属性不空时,就会调用 `predicates.NewServiceAffinityPredicate` 方法来创建一个处理服务亲和性的 `FitPredicate`, 随后被加载到全局的 `predicateFactory` 中生效。

最后, `genericScheduler.Schedule` 方法才是真正实现 Pod 调度的方法, 我们看看这段完整代码:

```

func (g *genericScheduler) Schedule(pod *api.Pod, minionLister algorithm.
MinionLister) (string, error) {
    minions, err := minionLister.List()
    if err != nil {
        return "", err
    }
    if len(minions.Items) == 0 {
        return "", ErrNoNodesAvailable
    }

    filteredNodes, failedPredicateMap, err := findNodesThatFit(pod, g.pods,
g.predicates, minions)
    if err != nil {
        return "", err
    }

    priorityList, err := PrioritizeNodes(pod, g.pods, g.prioritizers, algorithm.
FakeMinionLister(filteredNodes))
    if err != nil {
        return "", err
    }
    if len(priorityList) == 0 {
        return "", &FitError{
            Pod:      pod,

```

```

        FailedPredicates: failedPredicateMap,
    }
}

return g.selectHost(priorityList)
}

```

这段代码已经简单得不能再简单了，因为该干的活都已经被 `predicates` 与 `priorities` 干完了！架构之美，就在于程序逻辑分解得恰到好处，每个组件各司其职，从而化繁为简，使得主体流程清晰直观，犹如行云流水，一气呵成。

向谷歌大神们致敬！

### 6.4.3 设计总结

与之前的 Kubernetes API Server 和 Kubernetes Controller Manager 对比，Kubernetes Scheduler Server 的设计和代码显得更为“精妙”。项目中引入 `ratelimit` 组件来解决 Pod 调度的流控问题的做法，既大大简化了代码量，又体现了大神们的气度。

Kubernetes Scheduler Server 的一个关键设计目标是“插件化”，以方便 Cloud Provider 或者个人用户根据自己的需求进行定制，本节我们围绕其中最为关键的“FitPredicate 与 PriorityFunction”对其设计做一个总结。如图 6.7 所示，在 `plugin.go` 中采用了全局变量的 Map 变量记录了系统当前注册的 FitPredicate 与 PriorityFunction，其中 `fitPredicateMap` 和 `priorityFunctionMap` 分别存放 FitPredicateFactory 与 PriorityConfigFactory（包含了 PriorityFunctionFactory 的一个引用）中。可以看出，这里的设计采用了标准的工厂模式，`factory.PluginFactoryArgs` 这个数据结构可以认为是一个上下文环境变量，它提供给 PluginFactory 必要的的数据访问接口，比如获取一个 Node 的详细信息并获取一个 Pod 上的所有 Service 信息等，这些接口可以被某些具体的 FitPredicate 或 PriorityFunction 使用，以实现特定的功能，如图 6.7 所示的 `predicates.PodFitsPods` 和 `priorities.LeastRequestedPriority` 就分别使用了上述接口。

我们注意到 `PluginFactoryArgs` 的接口都是 Kubernetes 的资源访问接口，那么问题就来了，为何不直接用 Kubernetes RestClient API 访问呢？一个主要的原因是如果这样做，则增加了插件开发者开发和调测的难度，因为开发者需要再去学习和掌握 RestClient；另外一个原因是效率的问题，如果大家都采用框架提供的“标准方法”查询资源，那么框架可以实现很多优化，比较容易缓存；最后一个原因则与之前我们分析的“Assumed Pod”有关，即查询当前已经调度过的 Pod 列表是有其特殊性的，`PluginFactoryArgs` 中的 `PodLister` 方法就是引用了 `ConfigFactory` 的 `PodLister`。

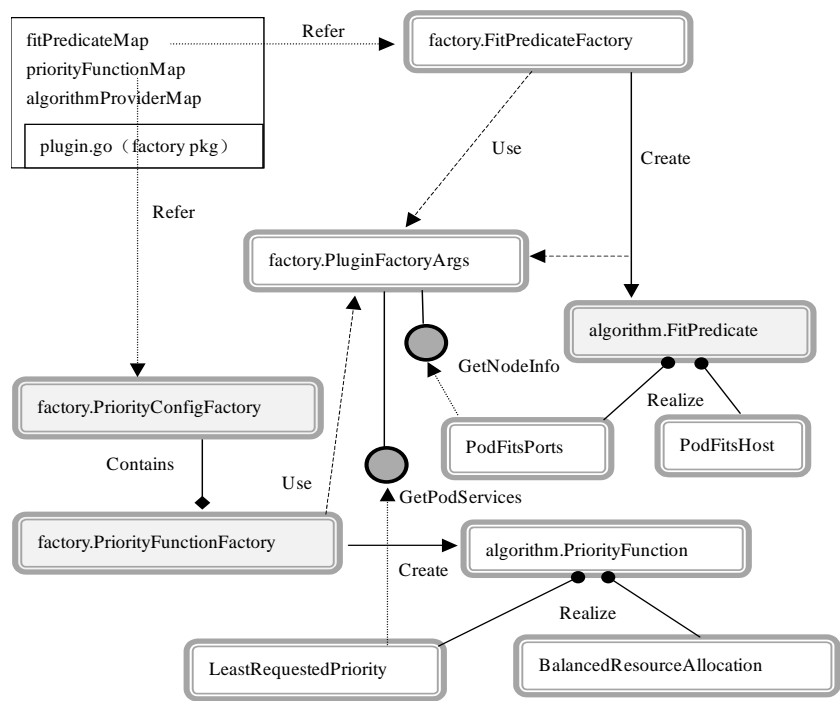


图 6.7 Kubernetes Scheduler Server 调度策略的相关设计示意图

`algorithmProviderMap` 这个全局变量则保存了一组命名的调度策略配置文件（`Algorithm ProviderConfig`），其实就是一组 `FitPredicate` 与 `PriorityFunction` 的集合，其定义如下：

```
type AlgorithmProviderConfig struct {
    FitPredicateKeys    util.StringSet
    PriorityFunctionKeys util.StringSet
}
```

它的作用是预配置和自定义调度规则，Kubernetes Scheduler Server 默认加载了一个名为“DefaultProvider”的调度策略配置，通过定义和加载不同的调度规则配置文件，我们可以改变默认的调度策略，比如我们可以定义两组规则文件：其中一个命名为“function\_test\_cfg”，面向功能测试，调度原则是尽量在最少的机器上调度 Pod 以节省资源；另外一个则命名为 performance\_test\_cfg”，面向性能测试，调度原则是尽可能使用更多的机器，以测试系统性能。

顺便提一下，笔者认为在 Kubernetes Scheduler Server 中关于 `PredicateArgument/Priority Argument` 的设计并不好，这里没有将 `Predicate` 的属性通用化，比如采用 key-value 这种模式，因此导致 Policy 文件格式与 `Predicate/Priority` 关联之间的强耦合性，增加了代码理解的困难性，之前分析的 Policy 文件中服务亲和性的 `Predicate` 的加载逻辑即反映了这个问题，笔者深信，未来版本中大神们会认真考虑重构问题。

至此，Master 节点上的进程的源码都已经分析完毕，我们发现这些进程所做的事情，归根到底就是两件事：Pod 调度+智能纠错，这也是为什么这些进程所在的节点被称为“Master”，因为它们高高在上，运筹帷幄。虽然“Master”从不深入底层微服私访，但也的确鞠躬尽瘁、日理万机，计算机的世界果然比我们人类的世界要单纯、高效很多，真心希望人工智能的发展不会让它们的世界也变得扑朔迷离。

## 6.5 kubelet 进程源码分析

kubelet 是运行在 Minion 节点上的重要守护进程，是工作在一线的重要“工人”，它才是负责“实例化”和“启动”一个具体的 Pod 的幕后主导，并且掌管着本节点上的 Pod 和容器的全生命周期过程，定时向 Master 汇报工作情况。此外，kubelet 进程也是一个“Server”进程，它默认监听 10250 端口，接收并执行远程（Master）发来的指令。

下面我们分别对其启动过程、关键代码分析及设计总结等方面进行深入分析和讲解。

### 6.5.1 进程启动过程

kubelet 进程的入口类源码位置如下：

`github.com/GoogleCloudPlatform/kubernetes/cmd/kubelet/kubelet.go`

入口 `main()` 函数的逻辑如下：

```
func main() {
    runtime.GOMAXPROCS(runtime.NumCPU())
    s := app.NewKubeletServer()
    s.AddFlags(pflag.CommandLine)
    util.InitFlags()
    util.InitLogs()
    defer util.FlushLogs()
    verflag.PrintAndExitIfRequested()
    if err := s.Run(pflag.CommandLine.Args()); err != nil {
        fmt.Fprintf(os.Stderr, "%v\n", err)
        os.Exit(1)
    }
}
```

我们已经是第 4 次“遇见”这样的代码风格了，代码的颜值匹配度高达 99%，这至少说明一点：谷歌在源码一致性方面做得很好，N 多人写的代码看起来就好像出自一个人之手。我们先来看看 KubeletServer 这个结构体所包括的属性吧，这些属性可以分为以下几组。

### 1) 基本配置

- ◎ **KubeConfig**: kubelet 默认配置文件路径。
- ◎ **Address、Port、ReadOnlyPort、CadvisorPort、HealthzPort、HealthzBindAddress**: 为 kubelet 绑定监听的地址，包括自身 Server 的地址，Cadvisor 绑定的地址，以及自身健康检查服务的绑定地址等。
- ◎ **RootDirectory、CertDirectory**: kubelet 默认的工作目录（/var/lib/kubelet），用于存放配置及 VM 卷等数据，CertDirectory 用于存放证书目录。

### 2) 管理 Pod 和容器相关的参数

- ◎ **PodInfraContainerImage**: Pod 的 infra 容器的镜像名称，谷歌被屏蔽时可以换成自己的私有仓库的镜像名。
- ◎ **CgroupRoot**: 可选项，创建 Pod 时所使用的顶层的 cgroup 名字（Root Cgroup）。
- ◎ **ContainerRuntime、DockerDaemonContainer、SystemContainer**: 这三个参数分别表示选择什么容器技术（Docker 或者 RKT）、Docker Daemon 容器的名字及可选的系统资源容器名称，用来将所有非 kernel 的、不在容器中的进程放入此容器中。

### 3) 同步和自动运维相关的参数

- ◎ **SyncFrequency、FileCheckFrequency、HTTPCheckFrequency**: Pod 容器同步周期、当前运行的容器实例分别与 Kubernetes 注册表中的信息、本地的 Pod 定义文件及以 HTTP 方式提供信息的数据源进行对比同步。
- ◎ **RegistryPullQPS、RegistryBurst**: 从注册表拉取待创建的 Pod 列表时的流控参数。
- ◎ **NodeStatusUpdateFrequency**: kubelet 多久汇报一次当前 Node 的状态。
- ◎ **ImageGCHighThresholdPercent、ImageGCLowThresholdPercent、LowDiskSpace ThresholdMB**: 分别是 Image 镜像占用磁盘空间的高低水位阈值及本机磁盘最小空闲容量，当可用容量低于这个容量时，所有新 Pod 的创建请求会被拒绝。
- ◎ **MaxContainerCount、MaxPerPodContainerCount**: 分别是 maximum-dead-containers 与 maximum-dead-containers-per-container，表示保留多少个死亡容器的实例在磁盘上，因为每个实例都会占用一定的磁盘，所以需要控制，默认是 MaxContainerCount 为 100，MaxPerPodContainerCount 为 2，即每个容器保留最多两个死亡实例，每个 Node 保留最多 100 个死亡实例。

只要分析一下上述 KubeletServer 结构体的关键属性，我们就可以得到这样一个推论：kubelet 进程的“工作量”还是很饱满的，一点都不比 Master 上的 API Server、Controll Manager、

Scheduer 做得少。

在继续下面的代码分析之前，我们先要理解这里的一个重要概念“Pod Source”，它是 kubelet 用于获取 Pod 定义和描述信息的一个“数据源”，kubelet 进程查询并监听 Pod Source 来获取属于自己所在节点的 Pod 列表，当前支持三种 Pod Source 类型。

- ◎ Config File: 本地配置文件作为 Pod 数据源。
- ◎ Http URL: Pod 数据源的内容通过一个 HTTP URL 方式获取。
- ◎ Kubernetes API Server: 默认方式，从 API Server 获取 Pod 数据源。

进程根据启动参数创建了 KubeletServer 以后，调用 KubeletServer 的 run 方法，进入启动流程，在流程的一开始首先设置了自身进程的 oom\_adj 参数（默认为-900），这是利用了 Linux 的 OOM 机制，当系统发生 OOM 时，oom\_adj 的值越小，越不容易被系统 Kill 掉。

```
if err := util.ApplyOomScoreAdj(0, s.OOMScoreAdj); err != nil {
    glog.Warning(err)
}
```

为什么在之前的 Master 节点进程上都没有见到这个调用，而在 kubelet 进程上却看到这段逻辑？答案很简单，因为 Master 节点不运行 Pod 和容器，主机资源通常是稳定和宽裕的，而 Minion 节点由于需要运行大量的 Pod 和容器，因此容易产生 OOM 问题，所以这里要确保“守护者”不会因此而被系统 Kill 掉。

由于 kubelet 会跟 API Server 打交道，所以接下来创建了一个 Rest Client 对象来访问 API Server。随后，启动进程构造了 cAdvisor 来监控本地的 Docker 容器，cAdvisor 具体的创建代码则位于 pkg/kubelet/cadvisor/cadvisor\_linux.go 里，引用了 github.com/google/cadvisor 这个同样属于谷歌开源的项目。

接着，初始化 CloudProvider，这是因为如果 Kubernetes 运行在某个运营商的 Cloud 环境中，则很多环境和资源都需要从 CloudProvider 中获取，比如在创建 Pod 的过程中可能需要知道某个 Node 的真实主机名。

虽然容器可以绑定宿主机的网络空间，但若不当使用会导致系统安全漏洞，所以 KubeletServer 中的 HostNetworkSources 的属性用来控制哪些 Pod 允许绑定宿主机的网络空间，默认是都禁止绑定。举例说明，比如设置 HostNetworkSources=api,http，则表明当一个 Pod 的定义来自 Kubernetes API Server 或者某个 HTTP URL 时，则允许此 Pod 绑定到宿主机的网络空间。下面这行代码即上述处理逻辑中的一小部分：

```
hostNetworkSources, err :=
kubelet.GetValidatedSources(strings.Split(s.HostNetworkSources, ","))
```

然后加载数字证书，如果没有提供证书和私钥，则默认创建一个自签名的 X509 证书并保

存到本地。下一步，创建一个 **Mounter** 对象，用来实现容器的文件系统挂载功能。

接下来的这段代码根据指定了 **DockerExecHandlerName** 参数的值，确定 **dockerExecHandler** 是采用 **Docker** 的 **exec** 命令还是 **nsenter** 来实现，默认采用了 **Docker** 的 **exec** 这种本地方式，**Docker** 从 1.3 版本开始提供了 **exec** 指令，为进入容器内部提供了更好的手段。

```
var dockerExecHandler dockertools.ExecHandler
switch s.DockerExecHandlerName {
case "native":
    dockerExecHandler = &dockertools.NativeExecHandler{}
case "nsenter":
    dockerExecHandler = &dockertools.NsenterExecHandler{}
default:
    log.Warningf("Unknown Docker exec handler %q; defaulting to native",
s.DockerExecHandlerName)
    dockerExecHandler = &dockertools.NativeExecHandler{}
}
```

运行至此，程序构造了一个 **KubeletConfig** 结构体，90%的变量与之前的 **KubeletServer** 一样，这让代码长度增加了 20 多行！定睛一看，源码上有 **TODO** 注释：“它应该可能被合并到 **KubeletServer** 里……”，目测注释是另外一个大神添加的，这让笔者陷入了深深的思考：难道谷歌的绩效考评系统中也有恶俗的代码行数考核指标？

**KubeletConfig** 创建好以后作为参数调用 **RunKubelet(&kcfg, nil)** 方法，程序运行到这里，才真正进入流程的核心步骤。下面这段代码表明 **kubelet** 会把自己的事件通知 **API Server**：

```
eventBroadcaster := record.NewBroadcaster()
kcfg.Recorder = eventBroadcaster.NewRecorder(api.EventSource{Component:
"kubelet", Host: kcfg.NodeName})
eventBroadcaster.StartLogging(glog.V(3).Infof)
if kcfg.KubeClient != nil {
    glog.V(4).Infof("Sending events to api server. ")
    eventBroadcaster.StartRecordingToSink(kcfg.KubeClient.Events(""))
} else {
    glog.Warning("No api server defined - no events will be sent to API server.
")
}
```

接着，启动进程进入关键函数 **createAndInitKubelet** 中，这里首先创建一个 **PodConfig** 对象，并根据启动参数中 **Pod Source** 参数是否提供，来创建相应类型的 **Pod Source** 对象，这些 **PodSource** 在各种协程中运行，拉取 **Pod** 信息并汇总输出到同一个 **Pod Channel** 中等待 **kubelet** 处理。创建 **PodConfig** 的具体代码如下：

```
func makePodSourceConfig(kc *KubeletConfig) *config.PodConfig {
    // source of all configuration
    cfg := config.NewPodConfig(config.PodConfigNotificationSnapshotAndUpdates,
```



```

kc.Recorder)

    // define file config source
    if kc.ConfigFile != "" {
        glog.Infof("Adding manifest file: %v", kc.ConfigFile)
        config.NewSourceFile(kc.ConfigFile, kc.NodeName, kc.FileCheckFrequency,
cfg.Channel(kubelet.FileSource))
    }

    // define url config source
    if kc.ManifestURL != "" {
        glog.Infof("Adding manifest url: %v", kc.ManifestURL)
        config.NewSourceURL(kc.ManifestURL, kc.NodeName, kc.HTTPCheckFrequency,
cfg.Channel(kubelet.HTTPSource))
    }
    if kc.KubeClient != nil {
        glog.Infof("Watching apiserver")
        config.NewSourceApiserver(kc.KubeClient, kc.NodeName, cfg.Channel
(kubelet.ApiserverSource))
    }
    return cfg
}

```

然后，创建一个 **kubelet** 并宣告它的诞生：

```

k, err = kubelet.NewMainKubelet(...)
k.BirthCry()

```

接着，触发 **kubelet** 开启垃圾回收协程以清理无用的容器和镜像，释放磁盘空间，下面是其代码片段：

```

// Starts garbage collection threads.
func (kl *Kubelet) StartGarbageCollection() {
    go util.Forever(func() {
        if err := kl.containerGC.GarbageCollect(); err != nil {
            glog.Errorf("Container garbage collection failed: %v", err)
        }
    }, time.Minute)

    go util.Forever(func() {
        if err := kl.imageManager.GarbageCollect(); err != nil {
            glog.Errorf("Image garbage collection failed: %v", err)
        }
    }, 5*time.Minute)
}

```

**createAndInitKubelet** 方法创建 **kubelet** 实例以后，返回到 **RunKubelet** 方法里，接下来调用 **startKubelet** 方法，此方法首先启动一个协程，让 **kubelet** 处理来自 **PodSource** 的 **Pod Update** 消息，

然后启动 **Kubelet Server**，下面是具体代码：

```
func startKubelet(k KubeletBootstrap, podCfg *config.PodConfig, kc *KubeletConfig) {
    // start the kubelet
    go util.Forever(func() { k.Run(podCfg.Updates()) }, 0)

    // start the kubelet server
    if kc.EnableServer {
        go util.Forever(func() {
            k.ListenAndServe(net.IP(kc.Address), kc.Port, kc.TLSOptions, kc.
EnableDebuggingHandlers)
        }, 0)
    }
    if kc.ReadOnlyPort > 0 {
        go util.Forever(func() {
            k.ListenAndServeReadOnly(net.IP(kc.Address), kc.ReadOnlyPort)
        }, 0)
    }
}
```

至此，**kubelet** 进程启动完毕。

## 6.5.2 关键代码分析

6.5.1 节里，我们分析了 **kubelet** 进程的启动流程，大致明白了 **kubelet** 的核心工作流程就是不断从 **Pod Source** 中获取与本节点相关的 **Pod**，然后开始“加工处理”，所以，我们先来分析 **Pod Source** 部分的代码。前面我们提到，**kubelet** 可以同时支持三类 **Pod Source**，为了能够将不同的 **Pod Source** “汇聚”到一起统一处理，谷歌特地设计了 **PodConfig** 这个对象，其代码如下：

```
type PodConfig struct {
    pods *podStorage
    mux  *config.Mux

    // the channel of denormalized changes passed to listeners
    updates chan kubelet.PodUpdate

    // contains the list of all configured sources
    sourcesLock sync.Mutex
    sources     util.StringSet
}
```

其中，**sources** 属性包括了当前加载的所有 **Pod Source** 类型，**sourcesLock** 是 **source** 的排他锁，在新增 **Pod Source** 的方法里使用它来避免共享冲突。

当 Pod 发生变动时，例如 Pod 创建、删除或者更新，相关的 Pod Source 就会产生对应的 PodUpdate 事件并推送到 Channel 上。为了能够统一处理来自多个 Source 的 Channel，谷歌设计了 config.Mux 这个“聚合器”，它负责监听多路 Channel，当接收到 Channel 发来的事件以后，交给 Merger 对象进行统一处理，Merger 对象最终把多路 Channel 发来的事件合并写入 updates 这个汇聚 Channel 里等待处理。

下面是 config.Mux 的结构体定义，其属性 sources 为一个 Channel Map，key 是对应的 Pod Source 的类型：

```
type Mux struct {
    // Invoked when an update is sent to a source.
    merger Merger
    // Sources and their lock.
    sourceLock sync.RWMutex
    // Maps source names to channels
    sources map[string]chan interface{}
}
```

我们继续深入分析 config.Mux 的工作过程，前面提到，kubelet 在启动过程中在 makePodSourceConfig 方法里创建了一个 PodConfig 对象，并且根据启动参数来决定要加载哪些类型的 Pod Source，在这个过程中调用了下述方法来创建一个对应的 Channel：

```
func (c *PodConfig) Channel(source string) chan<- interface{} {
    c.sourcesLock.Lock()
    defer c.sourcesLock.Unlock()
    c.sources.Insert(source)
    return c.mux.Channel(source)
}
```

而 Channel 具体的创建过程则在 config.Mux 里，Channel 创建完成后被加入 config.Mux 的 sources 里并且启动一个协程开始监听消息，代码如下：

```
func (m *Mux) Channel(source string) chan interface{} {
    if len(source) == 0 {
        panic("Channel given an empty name")
    }
    m.sourceLock.Lock()
    defer m.sourceLock.Unlock()
    channel, exists := m.sources[source]
    if exists {
        return channel
    }
    newChannel := make(chan interface{})
    m.sources[source] = newChannel
    go util.Forever(func() { m.listen(source, newChannel) }, 0)
    return newChannel
}
```

```
}
```

`config.Mux` 的上述 `listen` 方法很简单，就是监听新创建的 `Channel`，一旦发现 `Channel` 上有数据就交给 `Merger` 进行处理：

```
func (m *Mux) listen(source string, listenChannel <-chan interface{}) {
    for update := range listenChannel {
        m.merger.Merge(source, update)
    }
}
```

我们先来看看 `Pod Source` 是如何发送 `PodUpdate` 事件到自己所在的 `Channel` 上的，在 6.5.1 节中我们所见到的下面这段代码创建了一个 `Config File` 类型的 `Pod Source`：

```
// define file config source
if kc.ConfigFile != "" {
    glog.Infof("Adding manifest file: %v", kc.ConfigFile)
    config.NewSourceFile(kc.ConfigFile, kc.NodeName, kc.FileCheckFrequency,
cfg.Channel(kubelet.FileSource))
}
```

在 `NewSourceFile` 方法里启动了一个协程，每隔指定的时间（`kc.FileCheckFrequency`）就执行一次 `SourceFile` 的 `run` 方法，在 `run` 方法里所调用的主体逻辑是下面的函数：

```
func (s *sourceFile) extractFromPath() error {
    path := s.path
    statInfo, err := os.Stat(path)
    if err != nil {
        if !os.IsNotExist(err) {
            return err
        }
        // Emit an update with an empty PodList to allow FileSource to be marked
as seen
        s.updates <- kubelet.PodUpdate([]*api.Pod{}, kubelet.SET, kubelet.
FileSource)
        return fmt.Errorf("path does not exist, ignoring")
    }

    switch {
    case statInfo.Mode().IsDir():
        pods, err := s.extractFromDir(path)
        if err != nil {
            return err
        }
        s.updates <- kubelet.PodUpdate(pods, kubelet.SET, kubelet.FileSource)

    case statInfo.Mode().IsRegular():
        pod, err := s.extractFromFile(path)
        if err != nil {
```

```

        return err
    }
    s.updates <- kubelet.PodUpdate{[]*api.Pod{pod}, kubelet.SET, kubelet.
FileSource}

    default:
        return fmt.Errorf("path is not a directory or file")
    }

    return nil
}

```

看一眼上面的代码，我们就大致明白了 Config File 类型的 Pod Source 是如何工作的：它从指定的目录中加载多个 Pod 定义文件并转换为 Pod 列表或者加载单个 Pod 定义文件并转换为单个 Pod，然后生成对应的全量类型的 PodUpdate 事件并写入 Channel 中去。这里笔者也发现了代码命名的一个疏漏之处，SourceFile 的 updates 属性其实应该被命名为 update。其他两种 Pod Source 类型的代码解析就不在这里提及了。

接下来我们分析 Merger 对象，PodConfig 里的 Merger 对象其实是一个 config.podStorage 实例，它同时是 PodConfig 的 pods 属性的一个引用。podStorage 的源码位于 pkg/kubelet/config/config.go 里，其定义如下：

```

type podStorage struct {
    podLock sync.RWMutex
    // map of source name to pod name to pod reference
    pods map[string]map[string]*api.Pod
    mode PodConfigNotificationMode
    // ensures that updates are delivered in strict order
    // on the updates channel
    updateLock sync.Mutex
    updates    chan<- kubelet.PodUpdate
    // contains the set of all sources that have sent at least one SET
    sourcesSeenLock sync.Mutex
    sourcesSeen    util.StringSet
    // the EventRecorder to use
    recorder record.EventRecorder
}

```

我们看到 podStorage 的关键属性解释如下。

- (1) pods: 类型是 Map，存放每个 Pod Source 上拉过来的 Pod 数据，是 podStorage 当前保存“全量 Pod”的地方。
- (2) updates: 它就是 PodConfig 里的 updates 属性的一个引用。
- (3) mode: 表明 podStorage 的 Pod 事件通知模式，有以下几种。

- ◎ **PodConfigNotificationSnapshot**: 全量快照通知模式。
- ◎ **PodConfigNotificationSnapshotAndUpdates**: 全量快照+更新 Pod 通知模式（代码中创建 podStorage 实例时采用的模式）。
- ◎ **PodConfigNotificationIncremental**: 增量通知模式。

podStorage 实现的 Merge 接口的源码如下：

```
func (s *podStorage) Merge(source string, change interface{}) error {
    s.updateLock.Lock()
    defer s.updateLock.Unlock()
    adds, updates, deletes := s.merge(source, change)
    // deliver update notifications
    switch s.mode {
    case PodConfigNotificationSnapshotAndUpdates:
        if len(updates.Pods) > 0 {
            s.updates <- *updates
        }
        if len(deletes.Pods) > 0 || len(adds.Pods) > 0 {
            s.updates <- kubelet.PodUpdate{s.MergedState().([]*api.Pod), kubelet.SET, source}
        }
        //省略无关的 Case 逻辑
    }
    return nil
}
```

在上述 Merge 过程中，先调用内部函数 merge，将 Pod Source 的 Channel 上发来的 PodUpdate 事件分解为对应的新增、更新及删除等三类 PodUpdate 事件，然后判断是否有更新事件，如果有，则直接写入汇总的 Channel 中（podStorage.updates），然后调用 MergedState 函数复制一份 podStorage 的当前全量 Pod 列表，以此产生一个全量的 PodUpdate 事件并写入汇总的 Channel 中，从而实现了多 Pod Source Channel 的“汇聚”逻辑。

分析完 Merger 过程以后，我们接下来看看是什么对象，以及如何消费这个汇总的 Channel。在上一节提到，在 kubelet 进程启动的过程中调用了 startKubelet 方法，此方法首先启动一个协程，让 kubelet 处理来自 PodSource 的 Pod Update 消息，即下面这行代码：

```
go util.Forever(func() { k.Run(podCfg.Updates()) }, 0)
```

其中，PodConfig 的 Updates()方法返回了前面我们所说的汇总 Channel 变量的一个引用，下面是 kubelet 的 Run (updates <-chan PodUpdate)方法的代码：

```
func (kl *Kubelet) Run(updates <-chan PodUpdate) {
    if kl.logServer == nil {
        kl.logServer = http.StripPrefix("/logs/",
            http.FileServer(http.Dir("/var/log/")))
    }
}
```

```

    }
    if kl.kubeClient == nil {
        glog.Warning("No api server defined - no node status update will be sent. ")
    }
    // Move Kubelet to a container.
    if kl.resourceContainer != "" {
        err := util.RunInResourceContainer(kl.resourceContainer)
        if err != nil {
            glog.Warningf("Failed to move Kubelet to container %q: %v", kl.
resourceContainer, err)
        }
        glog.Infof("Running in container %q", kl.resourceContainer)
    }
    if err := kl.imageManager.Start(); err != nil {
        kl.recorder.Eventf(kl.nodeRef, "kubeletSetupFailed", "Failed to start
ImageManager %v", err)
        glog.Errorf("Failed to start ImageManager, images may not be garbage
collected: %v", err)
    }
    if err := kl.cadvisor.Start(); err != nil {
        kl.recorder.Eventf(kl.nodeRef, "kubeletSetupFailed", "Failed to start
CAdvisor %v", err)
        glog.Errorf("Failed to start CAdvisor, system may not be properly monitored:
%v", err)
    }
    if err := kl.containerManager.Start(); err != nil {
        kl.recorder.Eventf(kl.nodeRef, "kubeletSetupFailed", "Failed to start
ContainerManager %v", err)
        glog.Errorf("Failed to start ContainerManager, system may not be properly
isolated: %v", err)
    }
    if err := kl.oomWatcher.Start(kl.nodeRef); err != nil {
        kl.recorder.Eventf(kl.nodeRef, "kubeletSetupFailed", "Failed to start
OOM watcher %v", err)
        glog.Errorf("Failed to start OOM watching: %v", err)
    }
    go util.Until(kl.updateRuntimeUp, 5*time.Second, util.NeverStop)
    // Run the system oom watcher forever.
    kl.statusManager.Start()
    kl.syncLoop(updates, kl)
}

```

上述代码首先启动了一个 HTTP File Server 来远程获取本节点的系统日志，接下来根据启

动参数的设置来决定是否在指定的 Docker 容器中启动 kubelet 进程（如果成功，则将本进程转移到指定的容器中），然后分别启动 Image Manager（负责 Image GC）、cAdvisor（Docker 性能监控）、Container Manager（Container GC）、OOM Watcher（OOM 监测）、Status Manager（负责同步本节点上 Pod 的状态到 API Server 上）等组件，最后进入 syncLoop 方法中，无限循环调用下面的 syncLoopIteration 方法：

```
func (kl *Kubelet) syncLoopIteration(updates <-chan PodUpdate, handler
SyncHandler) {
    kl.syncLoopMonitor.Store(time.Now())
    if !kl.containerRuntimeUp() {
        time.Sleep(5 * time.Second)
        glog.Infof("Skipping pod synchronization, container runtime is not up. ")
        return
    }
    if !kl.doneNetworkConfigure() {
        time.Sleep(5 * time.Second)
        glog.Infof("Skipping pod synchronization, network is not configured")
        return
    }
    unsyncedPod := false
    podSyncTypes := make(map[types.UID]SyncPodType)
    select {
    case u, ok := <-updates:
        if !ok {
            glog.Errorf("Update channel is closed. Exiting the sync loop. ")
            return
        }
        kl.podManager.UpdatePods(u, podSyncTypes)
        unsyncedPod = true
        kl.syncLoopMonitor.Store(time.Now())
    case <-time.After(kl.resyncInterval):
        glog.V(4).Infof("Periodic sync")
    }
    start := time.Now()
    // If we already caught some update, try to wait for some short time
    // to possibly batch it with other incoming updates.
    for unsyncedPod {
        select {
        case u := <-updates:
            kl.podManager.UpdatePods(u, podSyncTypes)
            kl.syncLoopMonitor.Store(time.Now())
        case <-time.After(5 * time.Millisecond):
            // Break the for loop.
            unsyncedPod = false
        }
    }
}
```



```

pods, mirrorPods := kl.podManager.GetPodsAndMirrorMap()
kl.syncLoopMonitor.Store(time.Now())
if err := handler.SyncPods(pods, podSyncTypes, mirrorPods, start); err !=
nil {
    glog.Errorf("Couldn't sync containers: %v", err)
}
kl.syncLoopMonitor.Store(time.Now())
}

```

在上述代码中，如果从 Channel 中拉取到了 PodUpdate 事件，则先调用 podManager 的 UpdatePods 方法来确定此 PodUpdate 的同步类型，并将结果放入 podSyncTypes 这个 Map 中，同时为了提升处理效率，在代码中增加了持续循环拉取 PodUpdate 数据直到 Channel 为空为止（超时判断）的一段逻辑。在方法的最后，调用 SyncHandler 接口来完成 Pod 同步的具体逻辑，从而实现了 PodUpdate 事件的高效批处理模式。

SyncHandler 在这里就是 kubelet 实例本身，它的 SyncPods 方法比较长，其主要逻辑如下。

- ◎ 将传入的全量 Pod，与 statusManager 中当前保存的 Pod 集合进行对比，删除 statusManager 中当前已经不存在的 Pod（孤儿 Pod）。
- ◎ 调用 kubelet 的 admitPods 方法以过滤掉不适合本节点创建的 Pod。此方法首先过滤掉状态为 Failed 或者 Succeeded 的 Pod；接着过滤掉不适合本节点的 Pod，比如 Host Port 冲突、Node Label 的约束不匹配及 Node 的可用资源不足等情况；最后检查磁盘的使用情况，如果磁盘的可用空间不足，则过滤掉所有 Pod。
- ◎ 对上述过滤后的 Pod 集合中的每一个 Pod 调用 podWorkers 的 UpdatePod 方法，而此方法内部创建了一个 Pod 的 workUpdate 事件并发布到该 Pod 对应的一个 Work Channel 上（podWorkers.podWorkers）。
- ◎ 对于已经删除或不存在的 Pod，通知 podWorkers 删除相关联的 Work Channel（workUpdate）。
- ◎ 对比 Node 当前运行中的 Pod 及目标 Pod 列表，“杀掉”多余的 Pod，并且调用 Docker Runtime（Docker Deamon 进程）API，重新获取当前运行中的 Pod 列表信息。
- ◎ 清理“孤儿”Pod 所遗留的 PV 和磁盘目录。

要真正理解 Pod 是怎么在 Node 上“落地”的，还要继续深入分析上述第 3 步的代码。首先我们看看对 workUpdate 这个结构体的定义：

```

type workUpdate struct {
    pod *api.Pod
    // The mirror pod of pod; nil if it does not exist.
    mirrorPod *api.Pod
    // Function to call when the update is complete.
    updateCompleteFn func()
}

```

```

    updateType SyncPodType
}

```

其中的属性 `pod` 是当前要操作的 Pod 对象，`mirrorPod` 则是对应的镜像 Pod，下面是对它的解释：

对于每个来自非 API Server Pod Source 上的 Pod，kubelet 都在 API Server 上注册一个几乎“一模一样”的 Pod，这个 Pod 被称为 `mirrorPod`，这样一来，就将不同的 Pod Source 上的 Pod 都“统一”到了 kubelet 的注册表上，从而统一了 Pod 生命周期的管理流程。

`workUpdate` 的 `updateCompleteFn` 属性是一个回调函数，work 完成后会执行此回调函数，在上述第 3 步中，此函数用来计算该 work 的调度时延指标。

对于每个要同步的 Pod，`podWorkers` 会用一个长度为 1 的 Channel 来存放其对应的 `workUpdate`，而属性 `lastUndeliveredWorkUpdate` 则存放最近一个待安排执行的 `workUpdate`，这是因为一个 Pod 的前一个 `workUpdate` 正在执行时，可能会有一个新的 PodUpdate 事件需要处理。理解了这个过程后，再来看 `podWorkers` 的定义，就不难了：

```

type podWorkers struct {
    // Protects all per worker fields.
    podLock sync.Mutex
    podUpdates map[types.UID]chan workUpdate
    isWorking map[types.UID]bool
    lastUndeliveredWorkUpdate map[types.UID]workUpdate
    runtimeCache kubecontainer.RuntimeCache
    syncPodFn syncPodFnType
    recorder record.EventRecorder
}

```

下面这个函数就是第 3 步里产生 `workUpdate` 事件并放入到 `podWorkers` 的对应 Channel 的方法的源码：

```

func (p *podWorkers) UpdatePod(pod *api.Pod, mirrorPod *api.Pod, updateComplete
func()) {
    uid := pod.UID
    var podUpdates chan workUpdate
    var exists bool
    updateType := SyncPodUpdate
    p.podLock.Lock()
    defer p.podLock.Unlock()
    if podUpdates, exists = p.podUpdates[uid]; !exists {
        podUpdates = make(chan workUpdate, 1)
        p.podUpdates[uid] = podUpdates
        updateType = SyncPodCreate
        go func() {
            defer util.HandleCrash()
            p.managePodLoop(podUpdates)

```

```

    }()
}
if !p.isWorking[pod.UID] {
    p.isWorking[pod.UID] = true
    podUpdates <- workUpdate{
        pod:            pod,
        mirrorPod:       mirrorPod,
        updateCompleteFn: updateComplete,
        updateType:      updateType,
    }
} else {
    p.lastUndeliveredWorkUpdate[pod.UID] = workUpdate{
        pod:            pod,
        mirrorPod:       mirrorPod,
        updateCompleteFn: updateComplete,
        updateType:      updateType,
    }
}
}
}

```

上面的代码会调用 `podWorkers` 的 `managePodLoop` 方法来处理 `podUpdates` 队列，这里主要是获取必要的参数，最终处理又转手交给 `syncPodFn` 方法去处理。下面是 `managePodLoop` 的源码：

```

func (p *podWorkers) managePodLoop(podUpdates <-chan workUpdate) {
    var minRuntimeCacheTime time.Time
    for newWork := range podUpdates {
        func() {
            defer p.checkForUpdates(newWork.pod.UID, newWork.updateCompleteFn)
            if err := p.runtimeCache.ForceUpdateIfOlder(minRuntimeCacheTime); err != nil {
                glog.Errorf("Error updating the container runtime cache: %v", err)
                return
            }
            pods, err := p.runtimeCache.GetPods()
            if err != nil {
                glog.Errorf("Error getting pods while syncing pod: %v", err)
                return
            }
            err = p.syncPodFn(newWork.pod, newWork.mirrorPod,
                kubecontainer.Pods(pods).FindPodByID(newWork.pod.UID), newWork.
                updateType)
            if err != nil {
                glog.Errorf("Error syncing pod %s, skipping: %v", newWork.pod.UID, err)
                p.recorder.Eventf(newWork.pod, "failedSync", "Error syncing pod, skipping: %v",
                    err)
                return
            }
        }()
    }
}

```

```
        minRuntimeCacheTime = time.Now()
        newWork.updateCompleteFn()
    }()
}
```

追踪 `podWorkers` 的构造函数调用过程,可以发现 `syncPodFn` 函数其实就是 `kubelet` 的 `syncPod` 方法,这个方法的代码量有点儿多,主要逻辑如下。

(1) 根据系统配置中的权限控制,检查 Pod 是否有权在本节点运行,这些权限包括 Pod 是否有权使用 `HostNetwork` (还记得之前分析的代码么? 由 `Pod Source` 类型决定)、Pod 中的容器是否被授权以特权模式启动 (`privileged mode`) 等,如果未被授权,则删除当前运行中的旧版本的 Pod 实例并返回错误信息。

(2) 创建 Pod 相关的工作目录、PV 存放目录、Plugin 插件目录,这些目录都以 Pod 的 UID 为上一级目录。

(3) 如果 Pod 有 PV 定义,则针对每个 PV 执行目录的 `mount` 操作。

(4) 如果是 `SyncPodUpdate` 类型的 Pod,则从 `Docker Runtime` 的 API 接口查询获取 Pod 及相关容器的最新状态信息。

(5) 如果 Pod 有 `imagePullSecrets` 属性,则在 API Server 上获取对应的 `Secret`。

(6) 调用 `Container Runtime` 的 API 接口方法 `SyncPod`,实现 Pod “真正同步”的逻辑。

(7) 如果 Pod Source 不来自 API Server,则继续处理其关联的 `mirrorPod`。

◎ 如果 `mirrorPod` 跟当前 Pod 的定义不匹配,则它会被删除。

◎ 如果 `mirrorPod` 还不存在 (比如新创建的 Pod),则会在 API Server 上新建一个。

Kubernetes 中 `Container Runtime` 的默认实现是 `Dockers`,对应类是 `dockertools.DockerManager`,其源码位于 `kg/kubelet/dockertools/manager.go` 里,在上述 `kubelet.syncPod` 方法中所调用的 `DockerManager` 的 `SyncPod` 方法实现了下面的逻辑。

◎ 判断一个 Pod 实例的哪些组成部分需要重启: 包括 Pod 的 `infra` 容器是否发生变化 (如网络模式、Pod 里运行的各个容器的端口是否发生变化); Pod 里运行的容器是否发生变化; 用 `Probe` 检测容器的状态以确定容器是否异常等。

◎ 根据 Pod 实例重启结果的判断,如果需要重启 Pod 的 `infra` 容器,则先 Kill Pod 然后启动 Pod 的 `infra` 容器,设定好网络,最后启动 Pod 里的所有 `Container`; 否则就先 Kill 那些需要重启的 `Container`,然后重新启动它们。注意,如果是新创建的 Pod,则因为找不到 Node 上对应的 Pod 的 `infra` 容器,所以会被当作重启 Pod 的 `infra` 容器的逻辑来实现创建过程。

DockerManager 创建 Pod 的 infra 容器的逻辑在 createPodInfraContainer 方法里，大体逻辑如下。

- ◎ 如果 Pod 的网络不是 HostNetwork 模式，则搜集 Pod 所有容器的 Port 作为 infra 容器所要暴露的 Port 列表。
- ◎ 如果 infra 容器的 Image 目前不存在，则尝试拉取 Image。
- ◎ 创建 infra 的 Container 对象并且启动 runContainerInPod 方法。
- ◎ 如果容器定义有 Lifecycle，并且 PostStart 回调方法被设置了，就会触发此方法的调用，如果调用失败则 Kill 容器并返回。
- ◎ 创建一个软连接文件指向容器的日志文件，此软连接文件名包括 Pod 的名称、容器的名称及容器的 ID，这样的目的是让 Elasticsearch 这样的搜索技术容易索引和定位 Pod 日志。
- ◎ 如果此容器是 Pod infra 容器，则设置其 OOM 参数低于标准值，使得它比其他容器具备更强的“抗灾”能力。
- ◎ 修改 Docker 生成的容器的 resolv.conf 文件，增加 ndots 参数并默认设置为 5，这是因为 Kubernetes 默认假设的域名分割长度是 5，例如\_dns.\_udp.kube-dns.default.svc。

上述逻辑中所调用的 runContainerInPod 是 DockerManager 的核心方法之一，不管是创建 Pod 的 infra 容器还是 Pod 里的其他容器，都会通过此方法使得容器被创建和运行。以下是其主要逻辑。

- ◎ 生成 Container 必要的环境变量和参数，比如 ENV 环境变量、Volume Mounts 信息、端口映射信息、DNS 服务器信息、容器的日志目录、parent cgGroup 等。
- ◎ 调用 runContainer 方法完成 Docker Container 实例的创建过程，简单地说，就是完成 Docker create container 命令行所需的各种参数的构造过程，并通过程序来调用执行。
- ◎ 构造 HostConfig 对象，主要参数有目录映射、端口映射等、cgGroup 的设定等，简单地说，就是完成了 Docker start container 命令行所需的必要参数的构造过程，并通过程序来调用执行。

在上述逻辑中，runContainer 与 startContainer 的具体实现都是靠 DockerManager 中的 dockerClient 对象完成的，它实现了 DockerInterface 接口，dockerClient 的创建过程在 pkg/kubelet/dockertools/docker.go 里，下面是这段代码：

```
func ConnectToDockerOrDie(dockerEndpoint string) DockerInterface {
    if dockerEndpoint == "fake://" {
        return &FakeDockerClient{
            VersionInfo: docker.Env{"ApiVersion=1.18"},
        }
    }
}
```

```
    }
    client, err := docker.NewClient(getDockerEndpoint(dockerEndpoint))
    if err != nil {
        glog.Fatalf("Couldn't connect to docker: %v", err)
    }
    return client
}
```

这里的 `dockerEndpoint` 是本节点上的 Docker Deamon 进程的访问地址，默认是 `unix:///var/run/docker.sock`，在上述代码中使用了来自开源项目 <https://github.com/fsouza/go-dockerclient> 提供的 Docker Client，它也是 Go 语言实现的一个用 HTTP 访问 Docker Deamon 提供的标准 API 的客户端框架。

我们来看看 `dockerClient` 创建容器的具体代码（`CreateContainer`）：

```
func (c *Client) CreateContainer(opts CreateContainerOptions) (*Container, error) {
    path := "/containers/create? " + queryString(opts)
    body, status, err := c.do(
        "POST",
        path,
        doOptions{
            data: struct {
                *Config
                HostConfig *HostConfig `json: "HostConfig,omitempty" yaml:
"HostConfig,omitempty"`
            }{
                opts.Config,
                opts.HostConfig,
            },
        ),
    )
    if status == http.StatusNotFound {
        return nil, ErrNoSuchImage
    }
    if err != nil {
        return nil, err
    }
    var container Container
    err = json.Unmarshal(body, &container)
    if err != nil {
        return nil, err
    }
    container.Name = opts.Name
    return &container, nil
}
```

上述代码其实就是通过调用标准的 Docker Rest API 来实现功能的，我们进入 `docker.Client`

的 `do` 方法里可以看到更多详情，例如输入参数转换为 JSON 格式的数据、DockerAPI 版本检查及异常处理等逻辑，最有趣的是：在 `dockerEndpoint` 是 UNIX 套接字的情况下，会先建立套接字连接，然后在这个连接上创建 HTTP 连接。

至此，我们分析了 `kubelet` 创建和同步 Pod 实例的整个流程，简单总结如下。

- ◎ 汇总：先将多个 Pod Source 上过来的 PodUpdate 事件汇聚到一个总的 Channel 上去。
- ◎ 初审：分析并过滤掉不符合本节点的 PodUpdate 事件，对满足条件的 PodUpdate 则生成一个 workUpdate 事件，交给 podWorkers 处理。
- ◎ 接待：podWorkers 对每个 Pod 的 workUpdate 事件排队，并且负责更新 Cache 中的 Pod 状态，而把具体的任务转给 kubelet 去处理（`syncPod` 方法）。
- ◎ 终审：kubelet 对符合条件的 Pod 进一步审查，例如检查 Pod 是否有权在本节点运行，对符合审查的 Pod 开始着手准备工作，包括目录创建、PV 创建、Image 获取、处理 Mirror Pod 问题等，然后把“皮球”踢给了 DockerManager。
- ◎ 落地：任务抵达 DockerManager 之后，DockerManager 尽心尽责地分析每个 Pod 的情况，以决定这个 Pod 究竟是新建、完全重启还是部分更新的。给出分析结果以后，剩下的就是 `dockerClient` 的工作了。

好复杂的设计！原来非业务流程的代码理解起来也会如此折磨人，真心不知道谷歌当初是怎么设计和实现它的，目测国内 P8 水平的一帮大牛们天天加班到 9 点钟，也难以交付这样的 Code。

在继续下面的分析之前，留一个小小的思考给聪明的读者：Pod Source 上发来的 Pod 删除的事件，是在哪里处理的？

接下来我们继续分析 `kubelet` 进程的另外一个重要功能是如何实现的，即定期同步 Pod 状态信息到 API Server 上。先来看看 Pod 状态的数据结构定义：

```
type PodStatus struct {
    Phase      PodPhase      `json: "phase,omitempty"`
    Conditions []PodCondition `json: "conditions,omitempty"`
    Message string `json: "message,omitempty"`
    Reason string `json: "reason,omitempty"`
    HostIP string `json: "hostIP,omitempty"`
    PodIP string `json: "podIP,omitempty"`
    StartTime *util.Time `json: "startTime,omitempty"`
    ContainerStatuses []ContainerStatus
}

// PodStatusResult is a wrapper for PodStatus returned by kubelet that can be
encode/decoded
type PodStatusResult struct {
    TypeMeta `json: ",inline"`
```

```
ObjectMeta `json: "metadata,omitempty"`
Status PodStatus `json: "status,omitempty"`
}
```

Pod 的状态（Phase）有 5 种：运行中（PodRunning）、等待中（PodPending）、正常终止（PodSucceeded）、异常停止（PodFailed）及未知状态（PodUnknown），最后一种状态很可能是由于 Pod 所在主机的通信问题导致的。从上面的定义可以看到 Pod 的状态同时包括它里面运行的 Container 的状态，另外给出了导致当前状态的原因说明、Pod 的启动时间等信息。PodStatusResult 则是 Kubernetes API Server 提供的 Pod Status API 接口中用到的 Wrapper 类。

通过之前的代码研读，我们发现在 Kubernetes 中大量使用了 Channel 和协程机制来完成数据的高效传递和处理工作，在 kubelet 中更是大量使用了这一机制，实现 Pod Status 上报的 kubelet.statusManager 也是如此，它用一个 Map（podStatuses）保存了当前 kubelet 中所有 Pod 实例的当前状态，并且声明了一个 Channel（podStatusChannel）来存放 Pod 状态同步的更新请求（podStatuses），Pod 在本地实例化和同步的过程中会引发 Pod 状态的变化，这些变化被封装为 podStatusSyncRequest 放入 Channel 中，然后被异步上报到 API Server，这就是 statusManager 的运行机制。

下面是 statusManager 的 SetPodStatus 方法，先比较缓存的状态信息，如果状态发生变化，则触发 Pod 状态，生成 podStatusSyncRequest 并放到队列中等待上报：

```
func (s *statusManager) SetPodStatus(pod *api.Pod, status api.PodStatus) {
    podFullName := kubecontainer.GetPodFullName(pod)
    s.podStatusesLock.Lock()
    defer s.podStatusesLock.Unlock()
    oldStatus, found := s.podStatuses[podFullName]
    // ensure that the start time does not change across updates.
    if found && oldStatus.StartTime != nil {
        status.StartTime = oldStatus.StartTime
    }
    if status.StartTime.IsZero() {
        if pod.Status.StartTime.IsZero() {
            // the pod did not have a previously recorded value so set to now
            now := util.Now()
            status.StartTime = &now
        } else {
            status.StartTime = pod.Status.StartTime
        }
    }
    if !found || !isStatusEqual(&oldStatus, &status) {
        s.podStatuses[podFullName] = status
        s.podStatusChannel <- podStatusSyncRequest{pod, status}
    } else {
        glog.V(3).Infof("Ignoring same status for pod %q, status: %v", kubeletUtil.
```



```
FormatPodName(pod), status)
    }
}
```

下面是在 Pod 实例化的过程中, kubelet 过滤掉不合适本节点 Pod 所调用的上述方法的代码, 类似的调用还有不少:

```
func (kl *Kubelet) handleNotFittingPods(pods []*api.Pod) []*api.Pod {
    fitting, notFitting := checkHostPortConflicts(pods)
    for _, pod := range notFitting {
        reason := "HostPortConflict"
        kl.recorder.Eventf(pod, reason, "Cannot start the pod due to host port
conflict. ")
        kl.statusManager.SetPodStatus(pod, api.PodStatus{
            Phase:  api.PodFailed,
            Reason: reason,
            Message: "Pod cannot be started due to host port conflict"})
    }
    fitting, notFitting = kl.checkNodeSelectorMatching(fitting)
    for _, pod := range notFitting {
        reason := "NodeSelectorMismatching"
        kl.recorder.Eventf(pod, reason, "Cannot start the pod due to node selector
mismatch. ")
        kl.statusManager.SetPodStatus(pod, api.PodStatus{
            Phase:  api.PodFailed,
            Reason: reason,
            Message: "Pod cannot be started due to node selector mismatch"})
    }
    fitting, notFitting = kl.checkCapacityExceeded(fitting)
    for _, pod := range notFitting {
        reason := "CapacityExceeded"
        kl.recorder.Eventf(pod, reason, "Cannot start the pod due to exceeded
capacity. ")
        kl.statusManager.SetPodStatus(pod, api.PodStatus{
            Phase:  api.PodFailed,
            Reason: reason,
            Message: "Pod cannot be started due to exceeded capacity"})
    }
    return fitting
}
```

最后, 我们看看 statusManager 是怎么把 Channel 的数据上报到 API Server 上的, 这是通过 Start 方法开启一个协程无限循环执行 syncBatch 方法来实现的, 下面是 syncBatch 的代码:

```
func (s *statusManager) syncBatch() error {
    syncRequest := <-s.podStatusChannel
    pod := syncRequest.pod
    podFullName := kubecontainer.GetPodFullName(pod)
```

```
status := syncRequest.status

var err error
statusPod := &api.Pod{
    ObjectMeta: pod.ObjectMeta,
}
statusPod, err = s.kubeClient.Pods(statusPod.Namespace).Get(statusPod.Name)
if err == nil {
    statusPod.Status = status
    _, err = s.kubeClient.Pods(pod.Namespace).UpdateStatus(statusPod)
    // TODO: handle conflict as a retry, make that easier too.
    if err == nil {
        glog.V(3).Infof("Status for pod %q updated successfully", kubeletUtil.
FormatPodName(pod))
        return nil
    }
}
go s.DeletePodStatus(podFullName)
return fmt.Errorf("error updating status for pod %q: %v",
kubeletUtil.FormatPodName(pod), err)
}
```

这段代码首先从 Channel 中拉取一个 syncRequest，然后调用 API Server 接口来获取最新的 Pod 信息，如果成功，则继续调用 API Server 的 UpdateStatus 接口更新 Pod 状态，如果调用失败则删除缓存的 Pod 状态，这将触发 kubelet 重新计算 Pod 状态并再次尝试更新。

说完了 Pod 流程，我们接下来再一起深入分析 Kubernetes 中的容器探针（Probe）的实现机制。我们知道，容器正常不代表里面运行的业务进程能正常工作，比如程序还没初始化好，或者配置文件错误导致无法正常服务，还有诸如数据库连接爆满导致服务异常等各种意外情况都有可能发生，面对这类问题，cAdvisor 就束手无策了，所以 kubelet 引入了容器探针技术，容器探针按照作用划分为以下两种。

- ◎ **ReadinessProbe**: 用来探测容器中的用户服务进程是否处于“可服务状态”，此探针不会导致容器被停止或重启，而是导致此容器上的服务被标识为不可用，Kubernetes 不会发送请求到不可用的容器上，直到它们可用为止。
- ◎ **LivenessProbe**: 用来探测容器服务是否处于“存活状态”，如果服务当前被检测为 Dead，则会导致容器重启事件发生。

下面是探针相关的结构定义：

```
type Probe struct {
    Handler
    InitialDelaySeconds int64
    TimeoutSeconds int64
}
```

```

}
type Handler struct {
    // One and only one of the following should be specified.
    Exec *ExecAction
    HTTPGet *HTTPGetAction
    TCPSocket *TCPSocketAction
}

```

从上面的定义来看，探针可以通过执行容器中的一个命令、发起一个指向容器内部的 HTTP Get 请求或者 TCP 连接来确定容器内部是否正常工作。

上面的代码属于 API 包中的一部分，只是用来描述和存储容器上的探针定义，而真正的探针实现代码则位于 `pkg/kubelet/prober/prober.go` 里，下面是对 `prober.Probe` 的定义：

```

type Prober interface {
    Probe(pod *api.Pod, status api.PodStatus, container api.Container, containerID
string, createdAt int64) (probe.Result, error)
}

```

上述接口方法表示对一个 `Container` 发起探测并返回其结果。`prober.Probe` 的实现类为 `prober.prober`，其结构定义如下：

```

type prober struct {
    exec    execprobe.ExecProber
    http    httpprobe.HTTPProber
    tcp     tcpprobe.TCPProber
    runner  kubecontainer.ContainerCommandRunner
    readinessManager *kubecontainer.ReadinessManager
    refManager      *kubecontainer.RefManager
    recorder         record.EventRecorder
}

```

其中 `exec`、`http`、`tcp` 三个变量分别对应三种探测类型的“探头”，它们已经各自实现了相应的逻辑。比如下面这段代码是 HTTP 探头的核心逻辑，即连接一个 URL 发起 GET 请求：

```

func DoHTTPProbe(url *url.URL, client HTTPGetInterface) (probe.Result, string,
error) {
    res, err := client.Get(url.String())
    if err != nil {
        // Convert errors into failures to catch timeouts.
        return probe.Failure, err.Error(), nil
    }
    defer res.Body.Close()
    b, err := ioutil.ReadAll(res.Body)
    if err != nil {
        return probe.Failure, "", err
    }
    body := string(b)
    if res.StatusCode >= http.StatusOK && res.StatusCode < http.StatusBadRequest {

```

```
        glog.V(4).Infof("Probe succeeded for %s, Response: %v", url.String(),
*res)
        return probe.Success, body, nil
    }
    glog.V(4).Infof("Probe failed for %s, Response: %v", url.String(), *res)
    return probe.Failure, body, nil
}
```

`prober.prober` 中的 `runner` 则是 `exec` 探头的执行器，因为后者需要在被检测的容器中执行一个 `cmd` 命令：

```
func (p *prober) newExecInContainer(pod *api.Pod, container api.Container,
containerID string, cmd []string) exec.Cmd {
    return execInContainer(func() ([]byte, error) {
        return p.runner.RunInContainer(containerID, cmd)
    })
}
```

实际上 `p.runner` 就是之前我们分析过的 `DockerManager`，下面是 `RunInContainer` 的源码：

```
func (dm *DockerManager) RunInContainer(containerID string, cmd []string)
([]byte, error) {
    // If native exec support does not exist in the local docker daemon use nsinit.
    useNativeExec, err := dm.nativeExecSupportExists()
    if err != nil {
        return nil, err
    }
    if !useNativeExec {
        glog.V(2).Infof("Using nsinit to run the command %+v inside container
%s", cmd, containerID)
        return dm.runInContainerUsingNsinit(containerID, cmd)
    }
    glog.V(2).Infof("Using docker native exec to run cmd %+v inside container
%s", cmd, containerID)
    createOpts := docker.CreateExecOptions{
        Container:  containerID,
        Cmd:        cmd,
        AttachStdin: false,
        AttachStdout: true,
        AttachStderr: true,
        Tty:        false,
    }
    execObj, err := dm.client.CreateExec(createOpts)
    if err != nil {
        return nil, fmt.Errorf("failed to run in container - Exec setup failed
- %v", err)
    }
    var buf bytes.Buffer
```

```

startOpts := docker.StartExecOptions{
    Detach:      false,
    Tty:         false,
    OutputStream: &buf,
    ErrorStream:  &buf,
    RawTerminal:  false,
}
err = dm.client.StartExec(execObj.ID, startOpts)
if err != nil {
    glog.V(2).Infof("StartExec With error: %v", err)
    return nil, err
}
ticker := time.NewTicker(2 * time.Second)
defer ticker.Stop()
for {
    inspect, err2 := dm.client.InspectExec(execObj.ID)
    if err2 != nil {
        glog.V(2).Infof("InspectExec %s failed with error: %v", execObj.
ID, err2)
        return buf.Bytes(), err2
    }
    if !inspect.Running {
        if inspect.ExitCode != 0 {
            glog.V(2).Infof("InspectExec %s exit with result %v", execObj.
ID, inspect)
            err = &dockerExitError{inspect}
        }
        break
    }
    <-ticker.C
}

return buf.Bytes(), err
}

```

Docker 自 1.3 版本开始支持使用 Exec 指令（及 API 调用）在容器内执行一个命令，我们看看上述过程中使用的 `dm.client.CreateExec` 方法是如何实现的：

```

func (c *Client) CreateExec(opts CreateExecOptions) (*Exec, error) {
    path := fmt.Sprintf("/containers/%s/exec", opts.Container)
    body, status, err := c.do("POST", path, doOptions{data: opts})
    if status == http.StatusNotFound {
        return nil, &NoSuchContainer{ID: opts.Container}
    }
    if err != nil {
        return nil, err
    }
}

```

```
var exec Exec
err = json.Unmarshal(body, &exec)
if err != nil {
    return nil, err
}
return &exec, nil
}
```

我们看到，这是标准的 Docker API 的调用方式，与之前看到的创建容器的调用代码很相似。现在再回头看看 `prober.prober` 是怎么执行 `ReadinessProbe`/`LivenessProbe` 的检测逻辑的：

```
func (pb *prober) Probe(pod *api.Pod, status api.PodStatus, container api.Container, containerID string, createdAt int64) (probe.Result, error) {
    pb.probeReadiness(pod, status, container, containerID, createdAt)
    return pb.probeLiveness(pod, status, container, containerID, createdAt)
}
```

这段代码先调用容器的 `ReadinessProbe` 进行检测，并且在 `readinessManager` 组件中记录容器的 `Readiness` 状态，随后调用容器的 `LivenessProbe` 进行检测，并返回容器的状态，在检测过程中如果发现状态为失败或者异常状态，则会连续检测 3 次：

```
func (pb *prober) runProbeWithRetries(p *api.Probe, pod *api.Pod, status api.PodStatus, container api.Container, containerID string, retries int) (probe.Result, string, error) {
    var err error
    var result probe.Result
    var output string
    for i := 0; i < retries; i++ {
        result, output, err = pb.runProbe(p, pod, status, container, containerID)
        if result == probe.Success {
            return probe.Success, output, nil
        }
    }
    return result, output, err
}
```

比较意外的是 `prober.prober` 探针检测容器状态的方法目前只在一处被调用到，位于方法 `DockerManager.computePodContainerChanges` 里：

```
result, err := dm.prober.Probe(pod, podStatus, container, string(c.ID), c.Created)

if err != nil {
    // TODO(vmarmol): examine this logic.
    glog.V(2).Infof("probe no-error: %q", container.Name)
    containersToKeep[containerID] = index
    continue
}
if result == probe.Success {
```

```

        glog.V(4).Infof("probe success: %q", container.Name)
        containersToKeep[containerID] = index
        continue
    }
    glog.Infof("pod %q container %q is unhealthy (probe result: %v), it will
be killed and re-created. ", podFullName, container.Name, result)
    containersToStart[index] = empty{}
}

```

只有没有发生任何变化的 Pod 才会执行一次探针检测，若检测状态为失败，则会导致重启事件发生。

本节最后，我们再来简单分析 kubelet 中的 Kubelet Server 的实现机制，下面是 kubelet 进程启动过程中启动 Kubelet Server 的源码入口：

```

// start the kubelet server
if kc.EnableServer {
    go util.Forever(func() {
        k.ListenAndServe(net.IP(kc.Address), kc.Port, kc.TLSOptions, kc.
EnableDebuggingHandlers)
    }, 0)
}

```

在上述代码调用的过程中，创建了一个类型为 kubelet.Server 的 HTTP Server 并在本地监听：

```

handler := NewServer(host, enableDebuggingHandlers)
s := &http.Server{
    Addr:          net.JoinHostPort(address.String(), strconv.FormatUint
(uint64(port), 10)),
    Handler:       &handler,
    ReadTimeout:   5 * time.Minute,
    WriteTimeout:  5 * time.Minute,
    MaxHeaderBytes: 1 << 20,
}
if tlsOptions != nil {
    s.TLSConfig = tlsOptions.Config
    glog.Fatal(s.ListenAndServeTLS(tlsOptions.CertFile, tlsOptions.KeyFile))
} else {
    glog.Fatal(s.ListenAndServe())
}

```

在 kubelet.Server 的构造函数里加载如下 HTTP Handler：

```

func (s *Server) InstallDefaultHandlers() {
    healthz.InstallHandler(s.mux,
        healthz.PingHealthz,
        healthz.NamedCheck("docker", s.dockerHealthCheck),
        healthz.NamedCheck("hostname", s.hostnameHealthCheck),
        healthz.NamedCheck("syncloop", s.syncLoopHealthCheck),
    )
}

```

```

    )
    s.mux.HandleFunc("/pods", s.handlePods)
    s.mux.HandleFunc("/stats/", s.handleStats)
    s.mux.HandleFunc("/spec/", s.handleSpec)
}

```

上述 Handler 分为两组：首先是健康检查，包括 kubelet 进程自身的心跳检查、Docker 进程的健康检查、kubelet 所在主机名检测、Pod 同步的健康检查等；然后是获取当前节点上运行期信息的接口，例如获取当前节点上的 Pod 列表、统计信息等。下面是 `hostnameHealthCheck` 的实现逻辑，它检查 Pod 两次同步之间的时延，而这个时延则在之前提到的 kubelet 的 `syncLoopIteration` 方法中进行更新：

```

func (s *Server) syncLoopHealthCheck(req *http.Request) error {
    duration := s.host.ResyncInterval() * 2
    minDuration := time.Minute * 5
    if duration < minDuration {
        duration = minDuration
    }
    enterLoopTime := s.host.LatestLoopEntryTime()
    if !enterLoopTime.IsZero() && time.Now().After(enterLoopTime.Add(duration)) {
        return fmt.Errorf("Sync Loop took longer than expected. ")
    }
    return nil
}

```

`handlePods` 的 API 则从 kubelet 中获取当前“绑定”到本节点的所有 Pod 的信息并返回：

```

func (s *Server) handlePods(w http.ResponseWriter, req *http.Request) {
    pods := s.host.GetPods()
    data, err := encodePods(pods)
    if err != nil {
        s.error(w, err)
        return
    }
    w.Header().Add("Content-type", "application/json")
    w.Write(data)
}

```

如果 kubelet 运行在 Debug 模式，则加载更多的 HTTP Handler：

```

func (s *Server) InstallDebuggingHandlers() {
    s.mux.HandleFunc("/run/", s.handleRun)
    s.mux.HandleFunc("/exec/", s.handleExec)
    s.mux.HandleFunc("/portForward/", s.handlePortForward)

    s.mux.HandleFunc("/logs/", s.handleLogs)
    s.mux.HandleFunc("/containerLogs/", s.handleContainerLogs)
    s.mux.Handle("/metrics", prometheus.Handler())
}

```



```
// The /runningpods endpoint is used for testing only.
s.mux.HandleFunc("/runningpods", s.handleRunningPods)

s.mux.HandleFunc("/debug/pprof/", pprof.Index)
s.mux.HandleFunc("/debug/pprof/profile", pprof.Profile)
s.mux.HandleFunc("/debug/pprof/symbol", pprof.Symbol)
}
```

这些 HTTP Handler 的实现并不复杂，所以在这里就不再一一介绍了。

### 6.5.3 设计总结

在研读 kubelet 源码的过程中，你经常会有“山穷水尽疑无路，柳暗花明又一村”的感觉，是因为在它的设计中大量运用了 Channel 这种异步消息机制，加之为了测试的方便，又将很多重要的处理函数做成接口类，只有找到并分析这些接口的具体实现类，才能明白整个流程。这对于习惯了面向对象语言的程序员而言，有一种一夜回到解放前的感觉。

因为 kubelet 的功能比较多，所以我们在此仅以 Pod 同步的主流程为例，进行一个设计总结，图 6.8 是 kubelet 主流程相关的设计示意图，为了更加清晰地展示整个流程，我们特意将 kubelet Kernel、Docker System 与其他部分分离开来，并且省略了部分非核心对象和数据结构。

首先，config.PodConfig 创建一个或多个 Pod Source，在默认情况下创建的是 API source，它并没有创建新的数据结构，而是使用之前介绍的 cache.Reflector 结合 cache.UndeltaStore，从 Kubernetes API Server 上拉取 Pod 数据放入内部的 Channel 上，而内部的 Channel 收到 Pod 数据后会调用 podStorage 的 Merge 方法实现多个 Channel 数据的合并，产生 kubelet.PodUpdate 消息并写入 PodConfig 的汇总 Channel 上，随后 PodUpdate 消息进入 kubelet Kernel 中进行下一步处理。

kubelet.kubelet 的 syncLoop 方法监听 PodConfig 的汇总 Channel，过滤掉不合适的 PodUpdate 并把符合条件的放入 SyncPods 方法中，最终为每个符合条件的 Pod 产生一个 kubelet.workUpdate 事件并放入 podWorkers 的内部工作队列上，随后调用 podWorkers 的 managePodLoop 方法进行处理。podWorkers 在处理流程中调用了 DockerManager 的 SyncPod 方法，由此 DockerManager 接班，在进行了必要的 Pod 周边操作后，对于需要重启或者更新的容器，DockerManager 则交给 docker.Client 对象去执行具体的动作，后者通过调用 Dockers Engine 的 API Service 来实现具体功能。

在 Pod 同步的过程中会产生 Pod 状态的变更和同步问题，这些是交由 kubelet.statusManager 实现的，它在内部也采用了 Channel 的设计方式。

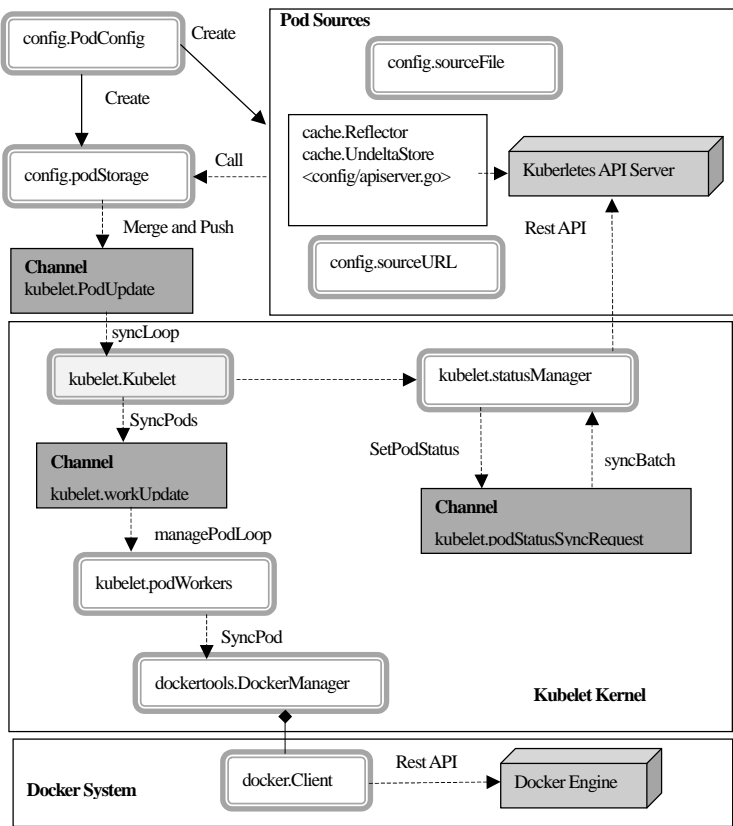


图 6.8 kubelet 主流程相关的设计示意图

## 6.6 kube-proxy 进程源码分析

kube-proxy 是运行在 Minion 节点上的另外一个重要的守护进程，你可以把它当作一个 HAProxy，它充当了 Kubernetes 中 Service 的负载均衡器和服务代理的角色。下面我们分别对其启动过程、关键代码分析及设计总结等方面进行深入分析和讲解。

### 6.6.1 进程启动过程

kube-proxy 进程的入口类源码码位置如下：

`github.com/GoogleCloudPlatform/kubernetes/cmd/kube-proxy/proxy.go`

入口 `main()` 函数的逻辑如下：

```
func main() {
    runtime.GOMAXPROCS(runtime.NumCPU())
    s := app.NewProxyServer()
    s.AddFlags(pflag.CommandLine)

    util.InitFlags()
    util.InitLogs()
    defer util.FlushLogs()

    verflag.PrintAndExitIfRequested()

    if err := s.Run(pflag.CommandLine.Args()); err != nil {
        fmt.Fprintf(os.Stderr, "%v\n", err)
        os.Exit(1)
    }
}
```

上述代码构造了一个 `ProxyServer`，然后调用它的 `Run` 方法启动运行。首先我们看看 `NewProxyServer` 的代码：

```
func NewProxyServer() *ProxyServer {
    return &ProxyServer{
        BindAddress:      util.IP(net.ParseIP("0.0.0.0")),
        HealthzPort:      10249,
        HealthzBindAddress: util.IP(net.ParseIP("127.0.0.1")),
        OOMScoreAdj:      -899,
        ResourceContainer: "/kube-proxy",
    }
}
```

在上述代码中，`ProxyServer` 绑定本地所有 IP（0.0.0.0）对外提供代理服务，而提供健康检查的 HTTP Server 则默认绑定本地的回环 IP，说明后者仅用于在本节点上访问，如果需要开发管理系统进行远程管理，则可以设置参数 `healthz-bind-address` 为 0.0.0.0 来达到目的。另外，从代码中看，`ProxyServer` 还有一个重要属性可以调整：`PortRange`（对应命令行参数为 `proxy-port-range`），它用来限定 `ProxyServer` 使用哪些本地端口作为代理端口，默认是随机选择。

`ProxyServer` 的 `Run` 方法流程如下。

- ◎ 设置本进程的 OOM 参数 `OOMScoreAdj`，保证系统 OOM 时，`kube-proxy` 不会首先被系统删除，这是因为 `kube-proxy` 与 `kubelet` 进程一样，比节点上的 Pod 进程更重要。
- ◎ 让自己的进程运行在指定的 Linux Container 中，这个 Container 的名字来自 `ProxyServer.ResourceContainer`，如上所述，默认为 `/kube-proxy`，比较重要的一点是这个 Container 具备所有设备的访问权。

- ◎ 创建 `ServiceConfig` 与 `EndpointsConfig`，它们与之前 `kubelet` 中的 `PodConfig` 的作用和实现机制有点像，分别负责监听和拉取 `API Server` 上 `Service` 与 `Service Endpoints` 的信息，并通知给注册到它们上的 `Listener` 接口进行处理。
- ◎ 创建一个 `round-robin` 轮询机制的 `load balancer` (`LoadBalancerRR`)，它用来实现 `Service` 的负载均衡转发逻辑，它也是前面创建的 `EndpointsConfig` 的一个 `Listener`。
- ◎ 创建一个 `Proxier`，它负责建立和维护 `Service` 的本地代理 `Socket`，它也是前面创建的 `ServiceConfig` 的一个 `Listener`。
- ◎ 创建一个 `config.SourceAPI`，并启动两个协程，通过 `Kubernetes Client` 来拉取 `Kubernetes API Server` 上的 `Service` 与 `Endpoint` 数据，然后分别写入之前定义的 `ServiceConfig` 与 `EndpointsConfig` 的 `Channel` 上，从而触发整个流程的驱动。
- ◎ 本地绑定健康检查的 `HTTP Server` 提供服务。
- ◎ 进入 `Proxier` 的 `SyncLoop` 方法里，该方法周期性地检查 `Iptables` 是否设置正常、服务的 `Portal` 是否正常开启，以及清除 `load balancer` 上的过期会话。

从启动流程看，`kube-proxy` 进程的参数比较少，它所做的事情也是比较单一的，没有 `kubelet` 进程那么复杂，在下一节我们会深入分析其关键代码。

## 6.6.2 关键代码分析

从上一节 `kube-proxy` 的启动流程来看，它跟 `kubelet` 有相似的地方，即都会从 `Kubernetes API Server` 拉取相关的资源数据并在本地节点上完成“深加工”，其拉取资源的做法，第一眼看上去与 `kubelet` 相似，但实际上有稍微不同的实现思路，这说明作者另有其人。

由于 `ServiceConfig` 与 `EndpointsConfig` 实现机制是完全一样的，只不过拉取的资源不同，所以我们这里仅对前者做深入分析。首先从 `ServiceConfig` 结构体开始：

```
type ServiceConfig struct {
    mux      *config.Mux
    bcaster  *config.Broadcaster
    store    *serviceStore
}
```

`ServiceConfig` 也使用了 `mux(config.Mux)`，它是一个多 `Channel` 的多路合并器，之前 `kubelet` 的 `PodConfig` 也用到了它。下面是 `ServiceConfig` 的构造函数：

```
func NewServiceConfig() *ServiceConfig {
    updates := make(chan struct{})
    store := &serviceStore{updates: updates, services:
make(map[string]map[types.NamespacedName]api.Service)}
```

```

mux := config.NewMux(store)
bcaster := config.NewBroadcaster()
go watchForUpdates(bcaster, store, updates)
return &ServiceConfig{mux, bcaster, store}
}

```

从上述代码来看，store 是 serviceStore 的一个实例。它作为 config.Mux 的 Merge 接口的实现，负责处理 config.Mux 的 Channel 上收到的 ServiceUpdate 消息并更新 store 的内部变量 services，后者是一个 Map，存放了最新同步到本地的 api.Service 资源，是 Service 的全量数据。下面是 Merge 方法的逻辑：

```

func (s *serviceStore) Merge(source string, change interface{}) error {
    s.serviceLock.Lock()
    services := s.services[source]
    if services == nil {
        services = make(map[types.NamespacedName]api.Service)
    }
    update := change.(ServiceUpdate)
    switch update.Op {
    case ADD:
        glog.V(4).Infof("Adding new service from source %s : %+v", source, update.
Services)
        for _, value := range update.Services {
            name := types.NamespacedName{value.Namespace, value.Name}
            services[name] = value
        }
    case REMOVE:
        glog.V(4).Infof("Removing a service %+v", update)
        for _, value := range update.Services {
            name := types.NamespacedName{value.Namespace, value.Name}
            delete(services, name)
        }
    case SET:
        glog.V(4).Infof("Setting services %+v", update)
        // Clear the old map entries by just creating a new map
        services = make(map[types.NamespacedName]api.Service)
        for _, value := range update.Services {
            name := types.NamespacedName{value.Namespace, value.Name}
            services[name] = value
        }
    default:
        glog.V(4).Infof("Received invalid update type: %v", update)
    }
    s.services[source] = services
    s.serviceLock.Unlock()
    if s.updates != nil {
        s.updates <- struct{}{}
    }
}

```

```
    }  
    return nil  
}
```

`serviceStore` 同时是 `config.Accessor` 接口的一个实现，`MergedState` 接口方法返回之前 Merge 最新的 `Service` 全量数据。

```
func (s *serviceStore) MergedState() interface{} {  
    s.serviceLock.RLock()  
    defer s.serviceLock.RUnlock()  
    services := make([]api.Service, 0)  
    for _, sourceServices := range s.services {  
        for _, value := range sourceServices {  
            services = append(services, value)  
        }  
    }  
    return services  
}
```

上述方法在哪里被用到了呢？就在之前提到的 `NewServiceConfig` 方法里：

```
go watchForUpdates(bcaster, store, updates)
```

一个协程监听 `serviceStore` 的 `updates(Channel)`，在收到事件以后就调用上述 `MergedState` 方法，将当前最新的 `Service` 数组通知注册到 `bcaster` 上的所有 `Listener` 进行处理。下面分别给出了 `watchForUpdates` 及 `Broadcaster` 的 `Notify` 方法的源码：

```
func watchForUpdates(bcaster *config.Broadcaster, accessor config.Accessor,  
updates <-chan struct{}) {  
    for true {  
        <-updates  
        bcaster.Notify(accessor.MergedState())  
    }  
}  
func (b *Broadcaster) Notify(instance interface{}) {  
    b.listenerLock.RLock()  
    listeners := b.listeners  
    b.listenerLock.RUnlock()  
    for _, listener := range listeners {  
        listener.OnUpdate(instance)  
    }  
}
```

上述逻辑的精巧设计之处在于，当 `ServiceConfig` 完成 Merge 调用后，为了及时通知 `Listener` 进行处理，就产生一个“空事件”并写入 `updates` 这个 `Channel` 中，另外监听此 `Channel` 的协程就及时得到通知，触发 `Listener` 的回调动作。`ServiceConfig` 这里注册的 `Listener` 是 `proxy.Proxier` 对象，我们以后会继续分析它的回调函数 `OnUpdate` 是如何使用 `Service` 数据的。

接下来，我们看看 `ServiceUpdate` 事件是怎么生成并传递到 `ServiceConfig` 的 `Channel` 上的。在 `kube-proxy` 启动流程中有调用 `config.NewSourceAPI` 函数，其内部生成了一个 `servicesReflector` 对象：

```
type servicesReflector struct {
    watcher      ServicesWatcher
    services     chan<- ServiceUpdate
    resourceVersion string
    waitDuration time.Duration
    reconnectDuration time.Duration
}
```

其中 `services` 这个 `Channel` 是用来写入 `ServiceUpdate` 事件的，它是 `ServiceConfig` 的 `Channel` (`source string`)方法所创建并返回的 `Channel`，它写入数据后就会被一个协程立即转发到 `ServiceConfig` 的 `Channel` 里。下面这段代码完整地揭示了上述逻辑：

```
func (c *ServiceConfig) Channel(source string) chan ServiceUpdate {
    ch := c.mux.Channel(source)
    serviceCh := make(chan ServiceUpdate)
    go func() {
        for update := range serviceCh {
            ch <- update
        }
        close(ch)
    }()
    return serviceCh
}
```

`servicesReflector` 中的 `watcher` 用来从 `API Server` 上拉取 `Service` 数据，它是 `client.Services` (`api.NamespaceAll`)返回的 `client.ServiceInterface` 实例对象的一个引用，属于标准的 `Kubernetes client` 包。在 `config.NewSourceAPI` 的方法里，启动了一个协程周期性地调用 `watcher` 的 `list` 与 `Watch` 方法获取数据，然后转换成 `ServiceUpdate` 事件，写入 `Channel` 中。下面是关键源码：

```
func (s *servicesReflector) run(resourceVersion *string) {
    if len(*resourceVersion) == 0 {
        services, err := s.watcher.List(labels.Everything())
        if err != nil {
            glog.Errorf("Unable to load services: %v", err)
            // TODO: reconcile with pkg/client/cache which doesn't use reflector.
            time.Sleep(wait.Jitter(s.waitDuration, 0.0))
            return
        }
        *resourceVersion = services.ResourceVersion
        // TODO: replace with code to update the
        s.services <- ServiceUpdate{Op: SET, Services: services.Items}
    }
}
```

```
    watcher, err := s.watcher.Watch(labels.Everything(), fields.Everything(),
    *resourceVersion)
    if err != nil {
        glog.Errorf("Unable to watch for services changes: %v", err)
        if !client.IsTimeout(err) {
            // Reset so that we do a fresh get request
            *resourceVersion = ""
        }
        time.Sleep(wait.Jitter(s.waitDuration, 0.0))
        return
    }
    defer watcher.Stop()
    ch := watcher.ResultChan()
    s.watchHandler(resourceVersion, ch, s.services)
}
```

在上面的代码中，初始时资源版本变量 `resourceVersion` 为空，于是会执行 `Service` 的全量拉取动作（`watcher.List`），之后 `Watch` 资源会开始发生变化（`watcher.Watch`）并将 `Watch` 的结果（一个 `Channel` 保持了 `Service` 的变动数据）也转换为对应的 `ServiceUpdate` 事件并写入 `Channel` 中。另外，当拉取数据的调用发生异常时，`resourceVersion` 恢复为空，导致重新进行全量资源的拉取动作。这种自修复能力的程序设计足以见证谷歌大神们的深厚编程功力；另外，笔者认为 `kube-proxy` 这里的 `ServiceConfig` 的设计实现思路和代码要比 `kubelet` 中的好一点，虽然两个作者都是顶尖高手。

接下来才开始进入本节的重点，即服务代理的实现机制分析。首先，我们从代码中的 `load balance` 组件说起。下面是 `kube-proxy` 中定义的 `Load Balancer` 接口：

```
type LoadBalancer interface {
    NextEndpoint(service ServicePortName, srcAddr net.Addr) (string, error)
    NewService(service ServicePortName, sessionAffinityType api.ServiceAffinity,
    stickyMaxAgeMinutes int) error
    CleanupStaleStickySessions(service ServicePortName)
}
```

`LoadBalancer` 有 3 个接口，其中 `NextEndpoint` 方法用于给访问指定 `Service` 的新客户端请求分配一个可用的 `Endpoint` 地址；`NewService` 用来添加一个新服务到负载均衡器上；`CleanupStaleStickySessions` 则用来清理过期的 `Session` 会话。目前 `kube-proxy` 只实现了一个基于 `round-robin` 算法的负载均衡器，它就是 `proxy.LoadBalancerRR` 组件。

`LoadBalancerRR` 采用了 `affinityState` 这个结构体来保存当前客户端的会话信息，然后在 `affinityPolicy` 里用一个 `Map` 来记录（属于某个 `Service` 的）所有活动的客户端会话，这是它实现 `Session` 亲和性的负载均衡调度的基础。

```
type affinityState struct {
    clientIP string
```



```

//clientProtocol api.Protocol //not yet used
//sessionCookie string //not yet used
endpoint string
lastUsed time.Time
}
type affinityPolicy struct {
    affinityType api.ServiceAffinity
    affinityMap map[string]*affinityState // map client IP -> affinity info
    ttlMinutes int
}

```

**balancerState** 用来记录一个 **Service** 的所有 **Endpoint**(数组)、当前所使用的 **Endpoint** 的 **index**，以及对应的所有活动的客户端会话 (**affinityPolicy**)。其定义如下：

```

type balancerState struct {
    endpoints []string // a list of "ip:port" style strings
    index int // current index into endpoints
    affinity affinityPolicy
}

```

有了上面的认识，再看 **LoadBalancerRR** 的构造函数就简单多了，它内部用一个 **map** 记录每个服务的 **balancerState** 状态，当然初始化时还是空的：

```

func NewLoadBalancerRR() *LoadBalancerRR {
    return &LoadBalancerRR{
        services: map[ServicePortName]*balancerState{},
    }
}

```

**LoadBalancerRR** 的 **NewService** 方法代码很简单，就是在它的 **services** 里增加一个记录项，用户端的会话超时时间 **ttlMinutes** 默认为 3h，下面是相关源码：

```

func (lb *LoadBalancerRR) NewService(svcPort ServicePortName, affinityType
api.ServiceAffinity, ttlMinutes int) error {
    lb.lock.Lock()
    defer lb.lock.Unlock()
    lb.newServiceInternal(svcPort, affinityType, ttlMinutes)
    return nil
}
func (lb *LoadBalancerRR) newServiceInternal(svcPort ServicePortName, affinityType
api.ServiceAffinity, ttlMinutes int) *balancerState {
    if ttlMinutes == 0 {
        ttlMinutes = 180
    }
    if _, exists := lb.services[svcPort]; !exists {
        lb.services[svcPort] = &balancerState{affinity:
        *newAffinityPolicy(affinityType, ttlMinutes)}
        glog.V(4).Infof("LoadBalancerRR service %q did not exist, created",
        svcPort)
    }
}

```

```

    } else if affinityType != "" {
        lb.services[svcPort].affinity.affinityType = affinityType
    }
    return lb.services[svcPort]
}

```

我们在前面提到过 `ServiceConfig` 同步并监听 API Server 上的 `api.Service` 的数据变化，然后调用 `Listener`（`proxy.Proxyier` 是 `ServiceConfig` 唯一注册的 `Listener`）的 `OnUpdate` 接口完成通知。而上述 `NewService` 就是在 `proxy.Proxyier` 的 `OnUpdate` 方法里被调用的，从而实现了 `Service` 自动添加到 `LoadBalancer` 的机制。

我们再来看 `LoadBalancerRR` 的 `NextEndpoint` 方法，它实现了经典的 `round-robin` 负载均衡算法。`NextEndpoint` 方法首先判断当前服务是否有保持会话（`sessionAffinity`）的要求，如果有，则看当前请求是否有连接可用：

```

if sessionAffinityEnabled {
    // Caution: don't shadow ipaddr
    var err error
    ipaddr, _, err = net.SplitHostPort(srcAddr.String())
    if err != nil {
        return "", fmt.Errorf("malformed source address %q: %v", srcAddr.
String(), err)
    }
    sessionAffinity, exists := state.affinity.affinityMap[ipaddr]
    if exists && int(time.Now().Sub(sessionAffinity.lastUsed).Minutes()) <
state.affinity.ttlMinutes {
        // Affinity wins.
        endpoint := sessionAffinity.endpoint
        sessionAffinity.lastUsed = time.Now()
        glog.V(4).Infof("NextEndpoint for service %q from IP %s with
sessionAffinity %+v: %s", svcPort, ipaddr, sessionAffinity, endpoint)
        return endpoint, nil
    }
}

```

如果服务无须会话保持、新建会话及会话过期，则采用 `round-robin` 算法得到下一个可用的服务端口，如果服务有会话保持需求，则保存当前的会话状态：

```

// Take the next endpoint.
endpoint := state.endpoints[state.index]
state.index = (state.index + 1) % len(state.endpoints)
if sessionAffinityEnabled {
    var affinity *affinityState
    affinity = state.affinity.affinityMap[ipaddr]
    if affinity == nil {
        affinity = new(affinityState) //&affinityState{ipaddr, "TCP", "",
endpoint, time.Now()}
    }
}

```

```

        state.affinity.affinityMap[ipaddr] = affinity
    }
    affinity.lastUsed = time.Now()
    affinity.endpoint = endpoint
    affinity.clientIP = ipaddr
    glog.V(4).Infof("Updated affinity key %s: %v", ipaddr, state.affinity.
affinityMap[ipaddr])
    }
    return endpoint, nil

```

接下来我们看看 Service 的 Endpoint 信息是如何添加到 LoadBalancerRR 上的？答案很简单，类似之前我们分析过的 ServiceConfig。kube-proxy 也设计了一个 EndpointsConfig 来拉取和监听 API Server 上的服务的 Endpoint 信息，并调用 LoadBalancerRR 的 OnUpdate 接口完成通知，在这个方法里，LoadBalancerRR 完成了服务访问端口的添加和同步逻辑。

我们先来看看 api.Endpoints 的定义：

```

type EndpointAddress struct {
    IP string
    TargetRef *ObjectReference
}
type EndpointPort struct {
    Name string
    Port int
    Protocol Protocol
}
type EndpointSubset struct {
    Addresses []EndpointAddress
    Ports     []EndpointPort
}
type Endpoints struct {
    TypeMeta `json: ",inline"`
    ObjectMeta `json: "metadata,omitempty"`
    Subsets []EndpointSubset
}

```

一个 EndpointAddress 与 EndpointPort 对象可以组成一个服务访问地址，而在 EndpointSubset 对象里则定义了两个单独的 EndpointAddress 与 EndpointPort 数组而不是“服务访问地址”的一个列表。初看这样的定义你可能会觉得很奇怪，为什么没有设计一个 Endpoint 结构？这里的深层次原因在于，Service 的 Endpoint 信息来源于两个独立的实体：Pod 与 Service，前者负责提供 IP 地址即 EndpointAddress，而后者负责提供 Port 即 EndpointPort。由于在一个 Pod 上可以运行多个 Service，而一个 Service 也通常跨越多个 Pod，于是就产生了一个“笛卡尔乘积”的 Endpoint 列表，这就是 EndpointSubset 的设计灵感。

举例说明，对于如下表示的 `EndpointSubset`：

```
{
  Addresses: [{ "ip": "10.10.1.1" }, { "ip": "10.10.2.2" }],
  Ports: [{ "name": "a", "port": 8675 }, { "name": "b", "port": 309 }]
}
```

会产生如下 `Endpoint` 列表：

```
a: [ 10.10.1.1:8675, 10.10.2.2:8675 ],
b: [ 10.10.1.1:309, 10.10.2.2:309 ]
```

`LoadBalancerRR` 的 `OnUpdate` 方法里循环对每个 `api.Endpoints` 进行处理，先把它转化为一个 `Map`，`Map` 的 `Key` 是 `EndpointPort` 的 `Name` 属性（代表一个 `Service` 的访问端口）；而 `Value` 则是 `hostPortPair` 的一个数组，`hostPortPair` 其实就是之前缺失的 `Endpoint` 结构体，包括一个 `IP` 地址与端口属性，即某个服务在一个 `Pod` 上的对应访问端口。

```
portsToEndpoints := map[string][]hostPortPair{}
for i := range svcEndpoints.Subsets {
  ss := &svcEndpoints.Subsets[i]
  for i := range ss.Ports {
    port := &ss.Ports[i]
    for i := range ss.Addresses {
      addr := &ss.Addresses[i]
      portsToEndpoints[port.Name] = append(portsToEndpoints[
[port.Name], hostPortPair{addr.IP, port.Port})
      // Ignore the protocol field - we'll get that from the Service
objects.
    }
  }
}
```

下一步，针对 `portsToEndpoints` 进行循环处理。对于每个记录，判断是否已经在 `services` 中存在，并做出相应的更新或跳过的逻辑，最后删除那些已经不在集合中的端口，完成整个同步逻辑。下面是相关代码：

```
for portname := range portsToEndpoints {
  svcPort := ServicePortName{types.NamespacedName{svcEndpoints.Namespace,
svcEndpoints.Name}, portname}
  state, exists := lb.services[svcPort]
  curEndpoints := []string{}
  if state != nil {
    curEndpoints = state.endpoints
  }
  newEndpoints := flattenValidEndpoints(portsToEndpoints[portname])

  if !exists || state == nil || len(curEndpoints) != len(newEndpoints)
|| !slicesEquiv(slice.CopyStrings(curEndpoints), newEndpoints) {
```

```

        glog.V(1).Infof("LoadBalancerRR: Setting endpoints for %s to %v",
svcPort, newEndpoints)
        lb.updateAffinityMap(svcPort, newEndpoints)
        // OnUpdate can be called without NewService being called externally.
        // To be safe we will call it here. A new service will only be created
        // if one does not already exist. The affinity will be updated
        // later, once NewService is called.
        state = lb.newServiceInternal(svcPort, api.ServiceAffinity(""), 0)
        state.endpoints = slice.ShuffleStrings(newEndpoints)

        // Reset the round-robin index.
        state.index = 0
    }
    registeredEndpoints[svcPort] = true
}
}
// Remove endpoints missing from the update.
for k := range lb.services {
    if _, exists := registeredEndpoints[k]; !exists {
        glog.V(2).Infof("LoadBalancerRR: Removing endpoints for %s", k)
        delete(lb.services, k)
    }
}
}

```

LoadBalancerRR 的代码总体来说还是比较简单的,它主要被 kube-proxy 中的关键组件 proxy.Proxier 所使用,后者用到的主要数据结构为 proxy.serviceInfo,它定义和保存了一个 Service 的代理过程中的必要参数和对象。下面是其定义:

```

type serviceInfo struct {
    portal          portal
    protocol        api.Protocol
    proxyPort       int
    socket          proxySocket
    timeout         time.Duration
    nodePort        int
    loadBalancerStatus api.LoadBalancerStatus
    sessionAffinityType api.ServiceAffinity
    stickyMaxAgeMinutes int
    // Deprecated, but required for back-compat (including e2e)
    deprecatedPublicIPs []string
}

```

serviceInfo 的各个属性解释如下。

- ◎ portal: 用于存放服务的 Portal 地址,即 Service 的 Cluster IP (VIP) 地址与端口。
- ◎ protocol: 服务的 TCP,目前是 TCP 与 UDP。

- ◎ **socket、proxyPort**: socket 是 Proxier 在本机上为该服务打开的代理 Socket; proxyPort 则是这个代理 Socket 的监听端口。
- ◎ **timeout**: 目前只用于 UDP 的 Service, 表明服务“链接”的超时时间。
- ◎ **nodePort**: 该服务定义的 NodePort。
- ◎ **loadBalancerStatus**: 在 Cloud 环境下, 如果存在由 Cloud 服务提供者提供的负载均衡器（软件或硬件）用作 Kubernetes Service 的负载均衡, 则这里存放这些负载均衡器的 IP 地址。
- ◎ **sessionAffinityType**: 该服务的负载均衡调度是否保持会话。
- ◎ **stickyMaxAgeMinutes**: 即前面说的 Session 过期时间。
- ◎ **deprecatedPublicIPs**: 已过期、废弃的服务的 Public IP 地址。

理解了 serviceInfo, 我们再来看 Proxier 的数据结构:

```
type Proxier struct {
    loadBalancer LoadBalancer
    mu            sync.Mutex // protects serviceMap
    serviceMap    map[ServicePortName]*serviceInfo
    portMapMutex  sync.Mutex
    portMap       map[portMapKey]ServicePortName
    numProxyLoops int32
    listenIP      net.IP
    iptables      iptables.Interface
    hostIP        net.IP
    proxyPorts    PortAllocator
}
```

Proxier 用一个 Map 维护了每个服务的 serviceInfo 信息, 同时为了快速查询和检测服务端口是否有冲突, 比如定义了两个一样端口的服务, 又设计了一个 portMap, 其 Key 为服务的端口信息 (portMapKey 由 port 和 protocol 组合而成), value 为 ServicePortName。Proxier 的 listenIP 为 Proxier 监听的本节点 IP, 它在这个 IP 上接收请求并做转发代理。由于每个服务的 proxySocket 在本节点监听的 Port 端口默认是系统随机分配的, 所以使用 PortAllocator 来分配这个端口。另外, Service 的 Portal 与 NodePort 是通过 Linux 防火墙机制来实现的, 因此这里引用了 Iptables 的组件完成相关操作。

要想理解 Proxier 中使用 Iptables 的方式, 首先我们要弄明白 Kubernetes 中 Service 访问的一些网络细节。先来看看图 6.9, 这是一个外部应用通过 NodePort (TCP: //NodeIP:NodePort) 来访问 Service 时的网络流量示意图。访问流量进入节点网卡 eth0 后, 到达 Iptables 的 PREROUTING 链, 通过 KUBE-NODEPORT-CONTAINER 这个 NAT 规则被转发到 kube-proxy 进程上该 Service 对应的 Proxy 端口, 然后由 kube-proxy 进程进行负载均衡并且将流量转发到

Service 所在 Container 的本地端口。

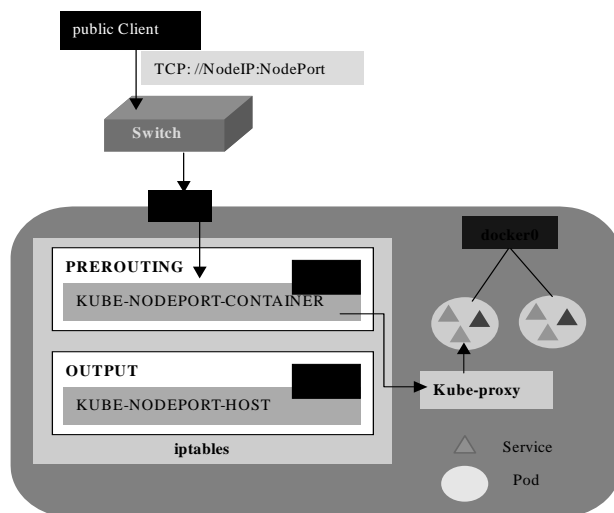


图 6.9 外部应用通过 NodePort 访问 Service 的网络流量示意图

根据 Iptables 的机制，本地进程发起的流量会经过 Iptables 的 OUTPUT 链，于是 kube-proxy 在这里也增加了相同作用的 NAT 规则：KUBE-NODEPORT-HOST。这样一来，如果本地容器内的进程以 NodePort 方式来访问 Service，则流量也会被转发到 kube-proxy 上，虽然以这种方式访问的情况比较少见。

服务之间通过 Service Portal 方式访问的流量转发机制与 NodePort 方式在本质上是相同的，也是通过 NAT，如图 6.10 所示。当 Service A 用 Service B 的 Portal 地址去访问时，流量经过 Iptables 的 OUTPUT 链经 NAT 规则 KUBE-PORTALS-HOST 的转换被转发到 kube-proxy 上，然后被转发给 Service B 所在的容器。

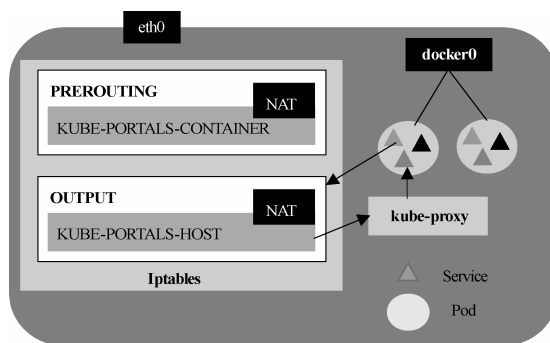


图 6.10 以 Service Portal 方式访问 Service 的流量示意图

Proxier 在创建 Iptables 的 PREROUTING 链中的 NAT 转发规则时，有一些特殊性，源码作者在代码中做了如下注释：

“这是一个复杂的问题。

如果 Proxy 的 Proxier.listenIP 设置为 0.0.0.0，即绑定到所有端口上，那么我们采用 REDIRECT 这种方式进行流量转发，因为这种情况下，返回的流量与进入的流量使用同一个网络端口，这就满足了 NAT 的规则。其他情况则采用 DNAT 转发流量，但 DNAT 到 127.0.0.1 时，流量会消失，这似乎是 Iptables 的一个众所周知的问题，所以这里不允许 Proxy 绑定到 localhost 上。”

现在再看下面这段代码就容易理解了，用来生成 KUBE-NODEPORT-CONTAINER 这条 NAT 规则：

```
func (proxier *Proxier) iptablesContainerNodePortArgs(nodePort int, protocol
api.Protocol, proxyIP net.IP, proxyPort int, serviceServicePortName) []string {
    args := iptablesCommonPortalArgs(nil, nodePort, protocol, service)
    if proxyIP.Equal(zeroIPv4) || proxyIP.Equal(zeroIPv6) {
        // TODO: Can we REDIRECT with IPv6?
        args = append(args, "-j", "REDIRECT", "--to-ports", fmt.Sprintf("%d",
proxyPort))
    } else {
        // TODO: Can we DNAT with IPv6?
        args = append(args, "-j", "DNAT", "--to-destination", net.JoinHostPort
(proxyIP.String(), strconv.Itoa(proxyPort)))
    }
    return args
}
```

弄明白 Proxier 中关于 Iptables 的事情之后，我们来研究和分析 Proxier 如何在 OnUpdate 方法里为每个 Service 建立起对应的 Proxy 并完成同步工作。首先，在 OnUpdate 方法里创建一个 map(activeServices) 来标识当前所有 alive 的 Service，key 为 ServicePortName，然后对 OnUpdate 参数里的 Service 数组进行循环，判断每个 Service 是否需要进行新建、变更或者删除操作，对于需要新建或者变更的 Service，先用 PortAllocator 获取一个新的未用的本地代理端口，然后调用 addServiceOnPort 方法创建一个 ProxySocket 用于实现此服务的代理，接着调用 openPortal 方法添加 Iptables 里的 NAT 映射规则，最后调用 LoadBalancer 的 NewService 方法把该服务添加到负载均衡器上。OnUpdate 方法的最后一段逻辑是处理已经被删除的 Service，对于每个要被删除的 Service，先删除 Iptables 中相关的 NAT 规则，然后关闭对应的 proxySocket，最后释放 ProxySocket 占用的监听端口并将该端口“还给”PortAllocator。



从上面的分析中，我们看到 `addServiceOnPort` 是 `Proxier` 的核心方法之一。下面是该方法的源码：

```
func (proxier *Proxier) addServiceOnPort(service ServicePortName, protocol
api.Protocol, proxyPort int, timeout time.Duration) (*serviceInfo, error) {
    sock, err := newProxySocket(protocol, proxier.listenIP, proxyPort)
    if err != nil {
        return nil, err
    }
    _, portStr, err := net.SplitHostPort(sock.Addr().String())
    if err != nil {
        sock.Close()
        return nil, err
    }
    portNum, err := strconv.Atoi(portStr)
    if err != nil {
        sock.Close()
        return nil, err
    }
    si := &serviceInfo{
        proxyPort:    portNum,
        protocol:     protocol,
        socket:       sock,
        timeout:      timeout,
        sessionAffinityType: api.ServiceAffinityNone, // default
        stickyMaxAgeMinutes: 180,                    // TODO: paramaterize this
in the API.
    }
    proxier.setServiceInfo(service, si)

    glog.V(2).Infof("Proxying for service %q on %s port %d", service, protocol,
portNum)
    go func(service ServicePortName, proxier *Proxier) {
        defer util.HandleCrash()
        atomic.AddInt32(&proxier.numProxyLoops, 1)
        sock.ProxyLoop(service, si, proxier)
        atomic.AddInt32(&proxier.numProxyLoops, -1)
    }(service, proxier)

    return si, nil
}
```

在上述代码中，先创建一个 `ProxySocket`，然后创建一个 `serviceInfo` 并添加到 `Proxier` 的 `serviceMap` 中，最后启动一个协程调用 `ProxySocket` 的 `ProxyLoop` 方法，使得 `ProxySocket` 进入

Listen 状态，开始接收并转发客户端请求。

kube-proxy 中的 ProxySocket 有两个实现，其中一个 `tcpProxySocket`，另外一个 `udpProxySocket`，二者的工作原理都一样，它们的工作流程就是为每个客户端 Socket 请求创建一个到 Service 的后端 Socket 连接，并且“打通”这两个 Socket，即把客户端 Socket 发来的数据“复制”到对应的后端 Socket 上，然后把后端 Socket 上服务响应的数据写入客户端 Socket 上去。

以 `tcpProxySocket` 为例，我们先看看它是如何完成 Service 后端连接创建过程的：

```
func tryConnect(service ServicePortName, srcAddr net.Addr, protocol string,
proxier *Proxier) (out net.Conn, err error) {
    for _, retryTimeout := range endpointDialTimeout {
        endpoint, err := proxier.loadBalancer.NextEndpoint(service, srcAddr)
        if err != nil {
            glog.Errorf("Couldn't find an endpoint for %s: %v", service, err)
            return nil, err
        }
        glog.V(3).Infof("Mapped service %q to endpoint %s", service, endpoint)
        outConn, err := net.DialTimeout(protocol, endpoint, retryTimeout*time.
Second)
        if err != nil {
            if isTooManyFDsError(err) {
                panic("Dial failed: " + err.Error())
            }
            glog.Errorf("Dial failed: %v", err)
            continue
        }
        return outConn, nil
    }
    return nil, fmt.Errorf("failed to connect to an endpoint. ")
}
```

在上述方法里，首先调用 `loadBalancer.NextEndpoint` 方法获取服务的下一个可用 Endpoint 地址，然后调用标准网络库中的方法建立至此地址的连接，如果连接失败，则会重新尝试，间隔时间指数增加（参见 `endpointDialTimeout` 的值）。

在后端 Service 的连接建立以后，`proxyTCP` 方法就会启动两个协程，通过调用 Go 标准库 `io` 里的 `Copy` 方法把输入流的数据写入输出流，从而完成前后端连接的数据转发功能。此外，`proxyTCP` 方法会阻塞，直到前后端两个连接的数据流都关闭（或结束）才会返回。下面是其源码：

```
func proxyTCP(in, out *net.TCPConn) {
    var wg sync.WaitGroup
    wg.Add(2)
```

```

glog.V(4).Infof("Creating proxy between %v <-> %v <-> %v <-> %v",
    in.RemoteAddr(), in.LocalAddr(), out.LocalAddr(), out.RemoteAddr())
go copyBytes("from backend", in, out, &wg)
go copyBytes("to backend", out, in, &wg)
wg.Wait()
in.Close()
out.Close()
}

```

这里我们留一个问题，kube-proxy 会在当前节点上为每个 Service 都建立一个代理么？不管本节点上是否有该 Service 对应的 Pod？

### 6.6.3 设计总结

从之前的启动流程和代码分析来看，kube-proxy 的设计和实现还是比较精巧和紧凑的，它的流程只有一个：从 Kubernetes API Server 上同步 Service 及其 Endpoint 信息，为每个 Service 建立一个本地代理以完成具备负载均衡能力的服务转发功能。图 6.11 给出了 kube-proxy 的总体设计示意图，为了清晰地表明整个业务流程和数据传递方向，这里省去了一些非关键的结构体和对象。app.ProxyServer 创建了一个 config.SourceAPI 的结构体，用于拉取 Kubernetes API Server 上的 Service 与 Endpoint 配置信息，分别由 config.servicesReflector 与 config.endpointsReflector 这两个对象来实现，它们各自通过相应的 Kubernetes Client API 来拉取数据并且生成对应的 Update 信息放入 Channel 中，最终 Channel 中的 Service 数据到达 proxy.Proxyer 上，proxy.Proxyer 为每个 Service 建立一个 proxySocket 实现服务代理并且在 Iptables 上创建相关的 NAT 规则，然后在 LoadBalancer 组件上开通该服务的负载均衡功能；而 Channel 中的 Endpoints 数据则被发送到 proxy.LoadBalancerRR 组件，用于给每个服务建立一个负载均衡的状态机，每个服务用 banlancerState 结构体来保存该服务可用的 Endpoint 地址及当前的会话状态 affinityPolicy，对于需要保存会话状态的服务，affinityPolicy 用一个 Map 来存储每个客户的会话状态 affinityState。

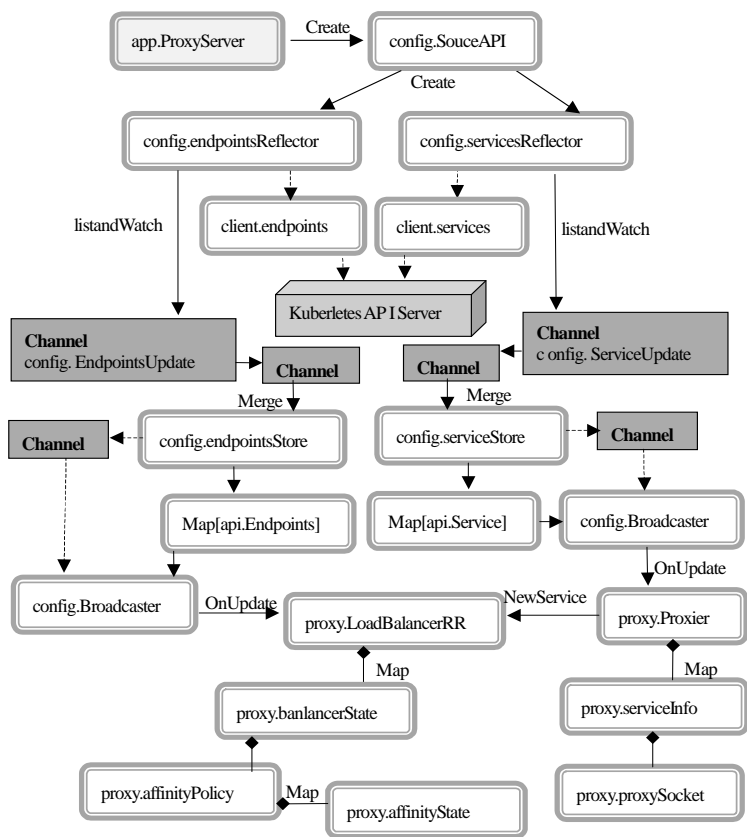


图 6.11 与 kubelet 总体相关的设计示意图

## 6.7 kubectI 进程源码分析

kubectI 与之前的 Kubernetes 进程不同，它不是一个后台运行的守护进程，而是 Kubernetes 提供的一个命令行工具（CLI），它提供了一组命令来操作 Kubernetes 集群。

kubectI 进程的入口类源码位置如下：

```
github.com/GoogleCloudPlatform/kubernetes/cmd/kubectI/kubectI.go
```

入口 main()函数的逻辑很简单：

```
func main() {
    runtime.GOMAXPROCS(runtime.NumCPU())
    cmd := cmd.NewKubectICommand(cmdutil.NewFactory(nil), os.Stdin, os.Stdout,
```

```
os.Stderr)
    if err := cmd.Execute(); err != nil {
        os.Exit(1)
    }
}
```

上述代码通过 `NewKubectlCommand` 方法创建了一个具体的 `Command` 命令并调用它的 `Execute` 方法执行，这是工厂模式结合命令模式的一个经典设计案例。从 `NewKubectlCommand` 的源码中可以看到，`kubectl` 的 CLI 命令框架使用了 GitHub 开源项目（<https://github.com/spf13/cobra>），下面是该框架中对 `Command` 的定义：

```
type Command struct {
    Use string // The one-line usage message.
    Short string // The short description shown in the 'help' output.
    Long string // The long message shown in the 'help <this-command>' output.
    Run func(cmd *Command, args []string) // Run runs the command.
}
```

实现一个具体 `Command` 就只要实现 `Command` 的 `Run` 函数即可，下面是其官方网站给出的一个 `Echo` 命令的例子：

```
var cmdEcho = &cobra.Command{
    Use: "echo [string to echo] ",
    Short: "Echo anything to the screen",
    Long: `echo is for echoing anything back.
Echo works a lot like print, except it has a child command.`,
    Run: func(cmd *cobra.Command, args []string) {
        fmt.Println("Print: " + strings.Join(args, " "))
    },
}
```

由于大多数 `kubectl` 的命令都需要访问 `Kubernetes API Server`，所以 `kubectl` 设计了一个类似命令的上下文环境的对象——`util.Factory` 供 `Command` 对象使用。

在接下来的几个章节中，我们对 `kubectl` 中的几个典型 `Command` 的源码逐一解读。

### 6.7.1 `kubectl create` 命令

`kubectl create` 命令通过调用 `Kubernetes API Server` 提供的 `Rest API` 来创建 `Kubernetes` 资源对象，例如 `Pod`、`Service`、`RC` 等，资源的描述信息来自 `-f` 指定的文件或者来自命令行的输入流。下面是创建 `create` 命令的相关源码：

```
func NewCmdCreate(f *cmdutil.Factory, out io.Writer) *cobra.Command {
    var filenames util.StringList
```

```

cmd := &cobra.Command{
    Use:     "create -f FILENAME",
    Short:   "Create a resource by filename or stdin",
    Long:    create_long,
    Example: create_example,
    Run: func(cmd *cobra.Command, args []string) {
        cmdutil.CheckErr(ValidateArgs(cmd, args))
        cmdutil.CheckErr(RunCreate(f, out, filenames))
    },
}
usage := "Filename, directory, or URL to file to use to create the resource"
kubectl.AddJsonFilenameFlag(cmd, &filenames, usage)
cmd.MarkFlagRequired("filename")
return cmd
}

```

`AddJsonFilenameFlag` 方法限制 `filename` 参数 (`-f`) 的文件名后缀只能是 `json`、`yaml` 或者 `yml` 中的一种，并且将参数值填充到 `filenames` 这个 `Set` 集合中，随后被 `Command` 的 `Run` 函数中的 `RunCreate` 方法所引用，后者就是 `kubectl create` 命令的核心逻辑所在。

`RunCreate` 方法使用到了 `resource.Builder` 对象，它是 `kubectl` 中的一处复杂设计，采用了 `Visitor` 的设计模式，`kubectl` 的很多命令都用到了它。`Builder` 的目标是根据命令行输入的资源相关的参数，创建针对性的 `Visitor` 对象来获取对应的资源，最后遍历相关的所有 `Visitor` 对象，触发用户指定的 `VisitorFun` 回调函数来处理每个具体的资源，最终完成资源对象的业务处理逻辑。由于涉及的资源参数有各种情况，所以导致 `Builder` 的代码很复杂。以下是 `Builder` 所能操作的各种资源参数：

- ◎ 通过输入流提供具体的资源描述；
- ◎ 通过本地文件内容或者 `HTTP URL` 的输出流来获取资源描述；
- ◎ 文件列表提供多个资源描述；
- ◎ 指定资源类型，通过查询 `Kubernetes API Server` 来获取相关类型的资源；
- ◎ 指定资源的 `selector` 条件如 `cluster-service=true`，查询 `Kubernetes API Server` 来获取相关的资源；
- ◎ 指定资源的 `namespace` 来查询符合条件的相关资源。

下面是 `resource.Builder` 的定义：

```

type Builder struct {
    mapper *Mapper
    errs []error
    paths []Visitor
}

```

```

    stream bool
    dir bool
    selector labels.Selector
    selectAll bool
    resources []string
    namespace string
    names []string
    resourceTuples []resourceTuple
    defaultNamespace bool
    requireNamespace bool
    flatten bool
    latest bool
    requireObject bool
    singleResourceType bool
    continueOnError bool
    schema validation.Schema
}

```

其实 **Builder** 很像一个 SQL 查询条件的生成器，里面包括了各种“查询”条件，在指定不同的查询条件时，会生成不同的 **Visitor** 接口来处理这些查询条件，最后遍历所有 **Visitor**，就得到最终的“查询结果”。**Builder** 返回的 **Result** 对象里也包括 **Visitor** 对象及可能的最终资源列表等信息，由于资源查询存在各种情况，所以 **Result** 也提供了多种方法，比如还包括了 **Watch** 资源变化的方法。

**RunCreate** 方法里先创建了一个 **Builder**，设置各种必要参数，然后调用 **Builder** 的 **Do** 方法，返回一个 **Result**，代码如下：

```

schema, err := f.Validator()
mapper, typer := f.Object()
r := resource.NewBuilder(mapper, typer, f.ClientMapperForCommand()).
    Schema(schema).
    ContinueOnError().
    NamespaceParam(cmdNamespace).DefaultNamespace().
    FilenameParam(enforceNamespace, filenames...).
    Flatten().
    Do()

```

其中，**schema** 对象用来校验资源描述是否正确，比如有没有缺少字段或者属性的类型错误等；**mapper** 对象用来完成从资源描述信息到资源对象的转换，用来在 **REST** 调用过程中完成数据转换；**FilenamParam** 是这里唯一指定 **Builder** 的资源参数的方法，即把命令行传入的 **filenames** 参数作为资源参数；**Flatten** 方法则告诉 **Builder**，这里的资源对象其实是一个数组，需要 **Builder** 构造一个 **FlattenListVisitor** 来遍历 **Visit** 数组中的每个资源项目；**Do** 方法则返回一个 **Rest** 对象，里面包括与资源相关的 **Visitor** 对象。

下面是 **NamespaceParam** 方法的源码，主要逻辑为调用 **Builder** 的 **Builder.Stdin**、**Builder.URL**

或 `Builder.Path` 方法来处理不同类型的资源参数，这些方法会生成对应的 `Visitor` 对象并加入 `Builder` 的 `Visitor` 数组里（`paths` 属性）。

```
func (b *Builder) FilenameParam(enforceNamespace bool, paths ...string) *Builder {
    for _, s := range paths {
        switch {
            case s == "-":
                b.Stdin()
            case strings.Index(s, "http:// ") == 0 || strings.Index(s, "https:// ") == 0:
                url, err := url.Parse(s)
                if err != nil {
                    b.errs = append(b.errs, fmt.Errorf("the URL passed to filename %q is not valid: %v", s, err))
                    continue
                }
                b.URL(url)
            default:
                b.Path(s)
            }
        }
    }
    if enforceNamespace {
        b.RequireNamespace()
    }
    return b
}
```

不管是标准输入流、URL，还是文件目录或者文件本身，这里处理资源的 `Visitor` 都是 `StreamVisitor` 这个实现（`FileVisitor` 与 `FileVisitorForSTDIN` 是 `StreamVisitor` 的一个 `Wrapper`）。下面是 `StreamVisitor` 的 `Visit` 接口代码：

```
func (v *StreamVisitor) Visit(fn VisitorFunc) error {
    d := yaml.NewYAMLOrJSONDecoder(v.Reader, 4096)
    for {
        ext := runtime.RawExtension{}
        if err := d.Decode(&ext); err != nil {
            if err == io.EOF {
                return nil
            }
            return err
        }
        ext.RawJSON = bytes.TrimSpace(ext.RawJSON)
        if len(ext.RawJSON) == 0 || bytes.Equal(ext.RawJSON, []byte("null")) {
            continue
        }
        if err := ValidateSchema(ext.RawJSON, v.Schema); err != nil {
            return err
        }
    }
}
```



```

    info, err := v.InfoForData(ext.RawJSON, v.Source)
    if err != nil {
        if v.IgnoreErrors {
            fmt.Fprintf(os.Stderr, "error: could not read an encoded object from
%s: %v\n", v.Source, err)
            glog.V(4).Infof("Unreadable: %s", string(ext.RawJSON))
            continue
        }
        return err
    }
    if err := fn(info); err != nil {
        return err
    }
}

```

在上述代码中，首先从输入流中解析具体的资源对象，然后创建一个 **Info** 结构体进行包装（转换后的资源对象存储在 **Info** 的 **Object** 属性中），最后再用这个 **Info** 对象作为参数调用回调函数 **VisitorFunc**，从而完成整个逻辑流程。下面是 **RunCreate** 方法里调用 **Builder** 的 **Visit** 方法触发 **Visitor** 执行时的源码，可以看到这里的 **VisitorFunc** 所做的事情是通过 **Rest Client** 发起 **Kubernetes API** 调用，把资源对象写入资源注册表里：

```

err = r.Visit(func(info *resource.Info) error {
    data, err := info.Mapping.Codec.Encode(info.Object)
    if err != nil {
        return cmdutil.AddSourceToErr("creating", info.Source, err)
    }
    obj, err := resource.NewHelper(info.Client, info.Mapping).Create(info.
Namespace, true, data)
    if err != nil {
        return cmdutil.AddSourceToErr("creating", info.Source, err)
    }
    count++
    info.Refresh(obj, true)
    printObjectSpecificMessage(info.Object, out)
    fmt.Fprintf(out, "%s/%s\n", info.Mapping.Resource, info.Name)
    return nil
})

```

## 6.7.2 rolling-update 命令

**kubectl rolling-update** 命令负责滚动更新（升级）**RC**（**ReplicationController**），下面是创建对应 **Command** 的源码：

```

func NewCmdRollingUpdate(f *cmdutil.Factory, out io.Writer) *cobra.Command {

```

```

cmd := &cobra.Command{
    Use: "rolling-update OLD_CONTROLLER_NAME ([NEW_CONTROLLER_NAME] -image
=NEW_CONTAINER_IMAGE | -f NEW_CONTROLLER_SPEC) ",
    // rollingupdate is deprecated.
    Aliases: []string{"rollingupdate"},
    Short:   "Perform a rolling update of the given ReplicationController. ",
    Long:    rollingUpdate_long,
    Example: rollingUpdate_example,
    Run: func(cmd *cobra.Command, args []string) {
        err := RunRollingUpdate(f, out, cmd, args)
        cmdutil.CheckErr(err)
    },
}

cmd.Flags().String("update-period", updatePeriod, `Time to wait between
updating pods. Valid time units are "ns", "us" (or "µs"), "ms", "s", "m", "h".`)

```

此处省去一些命令参数添加的非关键代码：

```

cmdutil.AddPrinterFlags(cmd)
return cmd
}

```

从上述代码中我们看到 `rolling-update` 命令的执行函数为 `RunRollingUpdate`，在分析这个函数之前，我们先了解下 `rolling-update` 执行过程中的一个关键逻辑。

`rolling update` 动作可能由于网络超时或者用户等得不耐烦等原因被中断，因此我们可能会重复执行一条 `rolling-update` 命令，目的只有一个，就是恢复之前的 `rolling update` 动作。为了实现这个目的，`rolling-update` 程序在执行过程中会在当前 `rolling-update` 的 RC 上增加一个 Annotation 标签——`kubectl.kubernetes.io/next-controller-id`，标签的值就是下一个要执行的新 RC 的名字。此外，对于 Image 升级这种更新方式，还会在 RC 的 Selector 上（`RC.Spec.Selector`）贴一个名为 `deploymentKey` 的 Label，Label 的值是 RC 的内容进行 Hash 计算后的值，相当于签名，这样就能很方便地比较 RC 里的 Image 名字（及其他信息）是否发生了变化。

`RunRollingUpdate` 执行逻辑的第 1 步：确定 New RC 对象及建立起 Old RC 到 New RC 的关联关系。下面我们以指定的 Image 参数进行 `rolling update` 的方式为例，看看代码是如何实现这段逻辑的。下面是相关源码：

```

if len(image) != 0 {
    keepOldName = len(args) == 1
    newName := findNewName(args, oldRc)
    if newRc, err = kubectl.LoadExistingNextReplicationController(client,
cmdNamespace, newName); err != nil {
        return err
    }
    if newRc != nil {
        fmt.Fprintf(out, "Found existing update in progress (%s), resuming.\n

```

```

n", newRc.Name)
    } else {
        newRc, err = kubect1.CreateNewControllerFromCurrentController(client,
cmdNamespace, oldName, newName, image, deploymentKey)
        if err != nil {
            return err
        }
    }
    // Update the existing replication controller with pointers to the 'next'
controller
    // and adding the <deploymentKey> label if necessary to distinguish it from
the 'next' controller.
    oldHash, err := api.HashObject(oldRc, client.Codec)
    if err != nil {
        return err
    }
    oldRc, err = kubect1.UpdateExistingReplicationController(client, oldRc,
cmdNamespace, newRc.Name, deploymentKey, oldHash, out)
    if err != nil {
        return err
    }
}
}

```

在代码里，`findNewName` 方法查询新 RC 的名字，如果在命令行参数中没有提供新 RC 的名字，则从 Old RC 中根据 `kubect1.kubernetes.io/next-controller-id` 这个 Annotation 标签找新 RC 的名字并返回，如果新 RC 存在则继续使用，否则调用 `CreateNewControllerFromCurrentController` 方法创建一个新 RC，在新 RC 的创建过程中设定 `deploymentKey` 的值为自己的 Hash 签名，方法源码如下：

```

func CreateNewControllerFromCurrentController(c *client.Client, namespace, oldName,
newName, image, deploymentKey string) (*api.ReplicationController, error) {
    // load the old RC into the "new" RC
    newRc, err := c.ReplicationControllers(namespace).Get(oldName)
    if err != nil {
        return nil, err
    }
    if len(newRc.Spec.Template.Spec.Containers) > 1 {
        // TODO: support multi-container image update.
        return nil, goerrors.New("Image update is not supported for multi-container
pods")
    }
    if len(newRc.Spec.Template.Spec.Containers) == 0 {
        return nil, goerrors.New(fmt.Sprintf("Pod has no containers! (%v) ",
newRc))
    }
    newRc.Spec.Template.Spec.Containers[0].Image = image
    newHash, err := api.HashObject(newRc, c.Codec)
}

```

```

    if err != nil {
        return nil, err
    }
    if len(newName) == 0 {
        newName = fmt.Sprintf("%s-%s", newRc.Name, newHash)
    }
    newRc.Name = newName
    newRc.Spec.Selector[deploymentKey] = newHash
    newRc.Spec.Template.Labels[deploymentKey] = newHash
    // Clear resource version after hashing so that identical updates get different
hashes.
    newRc.ResourceVersion = ""
    return newRc, nil
}

```

在 Image rolling update 的流程中确定新的 RC 以后，调用 `UpdateExistingReplicationController` 方法，将旧 RC 的 `kubectrl.kubernetes.io/next-controller-id` 设置为新 RC 的名字，并且判断旧 RC 是否需要设置或更新 `deploymentKey`，具体代码如下：

```

func UpdateExistingReplicationController(c client.Interface, oldRc *api.
ReplicationController, namespace, newName, deploymentKey, deploymentValue string,
out io.Writer) (*api.ReplicationController, error) {
    SetNextControllerAnnotation(oldRc, newName)
    if _, found := oldRc.Spec.Selector[deploymentKey]; !found {
        return AddDeploymentKeyToReplicationController(oldRc, c, deploymentKey,
deploymentValue, namespace, out)
    } else {
        // If we didn't need to update the controller for the deployment key, we still
need to write
        // the "next" controller.
        return c.ReplicationControllers(namespace).Update(oldRc)
    }
}

```

通过上面的逻辑，新 RC 被确定并且旧 RC 到新 RC 的关联关系也被建立好了，接下来如果 `dry-run` 参数为 `true`，则仅仅打印新旧 RC 的信息然后返回。如果是正常的 `rolling update` 动作，则创建一个 `kubectrl.RollingUpdater` 对象来执行具体任务，任务的参数则放在 `kubectrl.RollingUpdaterConfig` 中，相关源码如下：

```

updateCleanupPolicy := kubectrl.DeleteRollingUpdateCleanupPolicy
if keepOldName {
    updateCleanupPolicy = kubectrl.RenameRollingUpdateCleanupPolicy
}
config := &kubectrl.RollingUpdaterConfig{
    Out:      out,
    OldRc:    oldRc,
    NewRc:    newRc,

```

```

    UpdatePeriod: period,
    Interval:     interval,
    Timeout:      timeout,
    CleanupPolicy: updateCleanupPolicy,
}

```

其中 out 是输出流(屏幕输出); UpdatePeriod 是执行 rolling update 动作的间隔时间; Interval 与 Timeout 组合使用, 前者是每次拉取 polling controller 状态的间隔时间, 而后者则是对应的(HTTP REST 调用)超时时间。CleanupPolicy 确定升级结束后的善后策略, 比如 DeleteRollingUpdateCleanupPolicy 表示删除旧的 RC, 而 RenameRollingUpdateCleanupPolicy 则表示保持 RC 的名字不变(改变新 RC 的名字)。

RollingUpdater 的 Update 方法是 rolling update 的核心, 它以上述 config 对象作为参数, 其核心流程是每次让新 RC 的 Pod 副本数量加 1, 同时旧 RC 的 Pod 副本数量减 1, 直到新 RC 的 Pod 副本数量达到预期值同时旧 RC 的 Pod 副本数量变为零为止, 在这个过程中由于新旧 RC 的 Pod 副本数量一直在变动, 所以需要有一个地方记录最初不变的那个 Pod 副本数量, 这里就是 RC 的 Annotation 标签——kubectrl.kubernetes.io/desired-replicas。

下面这段源码就是“贴标签”的过程:

```

fmt.Fprintf(out, "Creating %s\n", newName)
    if newRc.ObjectMeta.Annotations == nil {
        newRc.ObjectMeta.Annotations = map[string]string{}
    }
    newRc.ObjectMeta.Annotations[desiredReplicasAnnotation] = fmt.Sprintf
("%d", desired)
    newRc.ObjectMeta.Annotations[sourceIdAnnotation] = sourceId
    newRc.Spec.Replicas = 0
    newRc, err = r.c.CreateReplicationController(r.ns, n

```

下面这段源码便是“江山代有才人出, 一代新人换旧人”的生动画面:

```

for newRc.Spec.Replicas < desired && oldRc.Spec.Replicas != 0 {
    newRc.Spec.Replicas += 1
    oldRc.Spec.Replicas -= 1
    fmt.Printf("At beginning of loop: %s replicas: %d, %s replicas: %d\n",
        oldName, oldRc.Spec.Replicas,
        newName, newRc.Spec.Replicas)
    fmt.Fprintf(out, "Updating %s replicas: %d, %s replicas: %d\n",
        oldName, oldRc.Spec.Replicas,
        newName, newRc.Spec.Replicas)
    newRc, err = r.scaleAndWait(newRc, retry, waitForReplicas)
    if err != nil {
        return err
    }
    time.Sleep(updatePeriod)
    oldRc, err = r.scaleAndWait(oldRc, retry, waitForReplicas)
}

```

```

    if err != nil {
        return err
    }
    fmt.Printf("At end of loop: %s replicas: %d, %s replicas: %d\n",
        oldName, oldRc.Spec.Replicas,
        newName, newRc.Spec.Replicas)
}
// delete remaining replicas on oldRc
if oldRc.Spec.Replicas != 0 {
    fmt.Fprintf(out, "Stopping %s replicas: %d -> %d\n",
        oldName, oldRc.Spec.Replicas, 0)
    oldRc.Spec.Replicas = 0
    oldRc, err = r.scaleAndWait(oldRc, retry, waitForReplicas)
    if err != nil {
        return err
    }
}
// add remaining replicas on newRc
if newRc.Spec.Replicas != desired {
    fmt.Fprintf(out, "Scaling %s replicas: %d -> %d\n",
        newName, newRc.Spec.Replicas, desired)
    newRc.Spec.Replicas = desired
    newRc, err = r.scaleAndWait(newRc, retry, waitForReplicas)
    if err != nil {
        return err
    }
}
}

```

上述方法里的 `scaleAndWait` 方法调用了 `kubectl.ReplicationControllerScaler` 的 `Scale` 方法，`Scale` 方法先通过 `Rest API` 调用 `Kubernetes API Server` 更新 `RC` 的 `Pod` 副本数量，然后循环拉取 `RC` 的信息，直到超时或者 `RC` 同步状态完成。下面是判断 `RC` 同步状态是否完成的函数，来自 `client` 包（`pkg/client/conditions.go`）。

```

func ControllerHasDesiredReplicas(c Interface, controller *api.ReplicationController)
wait.ConditionFunc {
    desiredGeneration := controller.Generation
    return func() (bool, error) {
        ctrl, err := c.ReplicationControllers(controller.Namespace).Get
        (controller.Name)
        if err != nil {
            return false, err
        }
        return ctrl.Status.ObservedGeneration >= desiredGeneration &&
        ctrl.Status.Replicas == ctrl.Spec.Replicas, nil
    }
}

```

rolling-update 是 kubectl 所有命令中最为复杂的一个，从它的功能和流程来看，完全可以被当作一个 Job 并放到 kube-controller-manager 上实现，客户端仅仅发起 Job 的创建及 Job 状态查看等命令即可，未来 Kubernetes 的版本是否会这样重构，我们拭目以待。

## 反侵权盗版声明

电子工业出版社依法对本作品享有专有出版权。任何未经权利人书面许可，复制、销售或通过信息网络传播本作品的行为；歪曲、篡改、剽窃本作品的行为，均违反《中华人民共和国著作权法》，其行为人应承担相应的民事责任和行政责任，构成犯罪的，将被依法追究刑事责任。

为了维护市场秩序，保护权利人的合法权益，我社将依法查处和打击侵权盗版的单位和个人。欢迎社会各界人士积极举报侵权盗版行为，本社将奖励举报有功人员，并保证举报人的信息不被泄露。

举报电话：（010）88254396；（010）88258888

传    真：（010）88254397

E-mail:    dbqq@phei.com.cn

通信地址：北京市万寿路 173 信箱

电子工业出版社总编办公室

邮    编：100036